CONTRIBUTIONS TO THE THEORY OF ROBUST REGRESSION

BY

ROBERT ALLEN WESLEY

TECHNICAL REPORT NO. 2
DECEMBER 1977

U.S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA
CONTRACT NO. DAAG29-76-G-0213

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **DEC 1977**	2. REPORT TYPE	3. DATES COVERED **00-00-1977 to 00-00-1977**
4. TITLE AND SUBTITLE **Contributions to the Theory of Robust Regression**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Stanford University,Department of Statistics,Stanford,CA,94305**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**90**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

CONTRIBUTIONS TO THE THEORY OF ROBUST REGRESSION

BY

ROBERT ALLEN WESLEY

TECHNICAL REPORT NO. 2
DECEMBER 1977

U.S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA
CONTRACT NO. DAAG29-76-G-0213

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ACKNOWLEDGEMENTS

There are many people to whom I am greatly indebted. I would first like to thank my parents for all of their encouragement and help. To Dr. Bentley I wish to extend my thanks for his friendship and for introducing me to statistics. Without the patience, advice, and good humor of my advisor, Professor M.V. Johns, this thesis would never have been completed. Lastly I wish to thank my wife, Margaret Nakamura Wesley, for her support, her ideas, and her love.

TABLE OF CONTENTS

LIST OF SYMBOLS AND NOTATION

For the sake of reference we record some of the more common notation and symbols used in the paper. For those symbols whose use in specific to this paper, we also record in parentheses the page(s) on which they are first used or where their definition may be found.

Standard Notation

In the background there is a probability space (Ω, A, P) on which all random variables are defined. The elements of Ω are denoted ω. For a random variable W, $E(W)$ refers to the expected value of W. Two random variables are of note: $N(0,1)$ is a normal random variable with mean 0 and variance 1; $U(a,b)$ is a random variable uniformly distributed on (a,b). If Z_n is a sequence of random variables, then $Z_n \overset{D}{\rightarrow} Z$ means that the Z_ns converge in distribution to the random variable Z. For a function $a(x)$, $a^-(x_o)$ is notation for $\lim_{x \uparrow x_o} a(x)$ and $a^+(x_o) = \lim_{x \downarrow x_o} a(x)$. w.p.1 (with probability one) and a.e. (almost everywhere) are equivalent notations. R and R^m are one-dimensional and m-dimensional Euclidean space respectively. If A is a set, $I_A(x)$ is the indicator function: $I_A(x) = 1$ if $x \in A$ and 0 if $x \notin A$. If $X_1, \ldots, X_n$ are random variables, the order statistic is $X_{(1)} \le X_{(2)} \le \cdot \cdot \cdot \le X_{(n)}$. Lastly, if A and B are two sets, $A \smallsetminus B$ is the set difference defined to equal $A \cap B^c$.

Symbols (Greek letters follow the Roman letters)

A: $a_N(\cdot)$ (8,10,17); $a_n(\cdot)$ (41,44); A (59); a(t) (61,67); $a^*(\cdot)$ (71).

B: B_f (18); B (18); B^o (18); B_J (18); b(N) (60); B_w (59); $B_{w'}$ (59); $b^*(\cdot)$ (71).

C: $\underset{\sim}{c}_i$ (1); c_{ij} (1); $\underset{\sim}{c}_i^*$ (1); c_i (2); $\bar{c}_j$ (19); C (10); C(u) (25).

D: $D_N(\cdot)$ (10,17).

E: e_i (1); e_i^β (12); $e_{(i)}^\beta$ (12); $\tilde{e}_j$ (59).

F: F (1); f (11); $\bar{F}_{\beta,N}$ (21); $\bar{F}_\beta$ (21); $F_{1n}, \ldots, F_{nn}$ (35); $\bar{F}$ (38,42); $\bar{F}_n$ (42); $\tilde{f}_N$ (59); f_N (59); $\tilde{F}_N^{-1}$ (61); F_N^{-1} (69).

G: G(y) (21,42).

H: $H(\cdot)$ (18); h (32); H_n (40); H_n^* (40); $\bar{H}_n$ (44).

I: I(f) (18); I_n (44); I_{1n}, I_{2n} (44); I (61); I^* (62).

J: J(u) (8,17); J_ξ (56).

K: K_β (22); K (42).

L: L_1, L_2 (47).

N: N (1).

P: $\underset{\sim}{p}_n$ (50,53).

Q: q (1).

S: S_N (20); S_n (35).

T: $t_1, \ldots, t_p$ (18,43); T, T_1, T_2 (64).

U: $U(\xi)$ (76).

V: $v(\xi)$ (57); $\tilde{v}_N(\xi)$ (73); $V(\xi)$, $V_N(\xi)$ (76).

W: $w(x)$ (59).

X: $X_{1n}, \ldots, X_{nn}$ (35); $X_{(i)}$ (35); X (50).

Y: Y_i (1,2,3,17); $\underset{\sim}{Y}$ (10); Y (50).

Z: Z_{kn} (38).

α: α (as a parameter) (1,2); α (as trimming proportion) (18);
α_0 (10).

β: $\underset{\sim}{\beta}$, $\underset{\sim}{\beta}^*$ (1); β (2); β_0 (10,17); $\tilde{\beta}_N$ (17); $\tilde{\underset{\sim}{\beta}}_N$ (58); $\underset{\sim}{\beta}_N^*$ (73).

ε: ε_1 (18,49).

ζ: ζ (62)

μ: $\mu(J,\overline{F}_{\beta,N})$ (23); $\mu_N(\beta)$ (23); $\mu(J,\overline{F}_\beta)$ (23); $\mu(\beta)$ (23);
$\mu(J,F)$ (36).

ξ: ξ_1 (61); ξ_0 (73); Ξ (73).

σ: Σ (19); $\sigma^2(J,\overline{F}_\beta,K)$ (22); σ_β^2 (22).

ϕ: ϕ_f (11).

CHAPTER 1. INTRODUCTION

1.1 Description of the Problem

The general problem we will be concerned with in this paper is that of estimating the regression parameters $\alpha, \beta_1, \ldots, \beta_q$ in the general regression model

$$(1.1) \qquad Y_i = \alpha + \underset{\sim}{c}_i^T \underset{\sim}{\beta} + e_i \qquad i = 1,2,\ldots,N,$$

where Y_i is the i^{th} observation on the dependent variable, α is the intercept parameter, $\underset{\sim}{\beta}$ is a column vector of slope parameters with $\underset{\sim}{\beta}^T = (\beta_1, \ldots, \beta_q)$ (T denotes transpose), $\underset{\sim}{c}_i^T = (c_{i1}, \ldots, c_{iq})$ is the vector of regression constants associated with the i^{th} observation, and e_i is the random error associated with the i^{th} observation. A formulation equivalent to (1.1) which is sometimes more convenient is

$$(1.2) \qquad Y_i = \underset{\sim}{c}_i^{*T} \underset{\sim}{\beta}^* + e_i \quad ,$$

where now $\underset{\sim}{c}_i^{*T} = (1, c_{i1}, \ldots, c_{iq})$ and $\underset{\sim}{\beta}^{*T} = (\alpha, \beta_1, \ldots, \beta_q)$. Throughout we will assume that the $\{e_i : i=1,\ldots,N\}$ are independent, identically distributed (iid) random variables (rv^s) with cumulative distribution function (cdf) $F(x)$, which is symmetric but possibly far from the normal distribution.

More specifically, throughout most of the paper we will consider

the simpler problem of estimating α, β in the simple linear regression case, for which the model (1.1) becomes

$$Y_i = \alpha+\beta c_i+e_i \quad ; \tag{1.3}$$

now β and c_i are scalars. Besides simplicity, another reason for considering (1.3) was pointed out by Huber in [15] :

> "Note that the simple straight line regression problem is basic; if we know how to treat this, we can in principle attack the general regression problem by considering one parameter at a time, keeping the others fixed at trial values."

The main feature of note in models (1.1) and (1.3) is that we do not assume that the random errors $\{e_i\}$ are normally distributed. Indeed we will not assume we know the form of the distribution function F. In this context the problem of estimating the regression parameters becomes a problem of robust estimation and shares many features with the problem of robust estimation of a location parameter, which has been considered extensively in the last decade (cf. [3] , [11] , [12] , [13] , [14] , [16] , [17] , [19]). Here we use the word "robust" to refer to statistical procedures good for a broad class of possible underlying models. Asymptotically this can be viewed as demanding high absolute (asymptotic) efficiencies for all suitably smooth shapes. (For this and other approaches to "robustness", consult [12] and [14].)

In the location problem the model in which we are interested is

$$Y_i = \theta + e_i \qquad , \tag{1.4}$$

where θ is the unknown location parameter, Y_i is the observation, and the $\{e_i\}$ are iid random variables with common cdf F assumed to be symmetric around 0. The simple linear regression model of (1.3) is one of the simplest gneralizations of this problem; both the regression models (1.3) and (1.1) include the location model as a special case and are important in practical applications.

For the location problem three different classes of estimators of θ have been proposed: L estimators, M estimators, and R estimators. Briefly summarizing: the L estimators are linear combinations of the ordered observations $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}$; the M estimators are analogues of maximum likelihood estimators, with the estimator $\hat{\theta}$ of θ satisfying

$$\sum_{i=1}^{N} \rho(Y_i - \hat{\theta}) = \text{minimum}, \tag{1.5}$$

or the equation

$$\sum_{i=1}^{N} \psi(Y_i - \hat{\theta}) = 0, \tag{1.6}$$

where ρ is usually a convex function and $\psi = \rho'$; R estimators, such

as the well-known Hodges-Lehmann estimator (cf. [11]), are derived from rank tests. For the problem of estimating regression parameters, each of these three classes of estimators has been generalized. In the next section we will consider these generalizations after briefly reviewing the classical technique of estimation for location and regression problems -- the method of least squares -- and some of its history.

1.2 Techniques for Regression Estimation

1.2.1 In the history of statistics, the problem of estimating regression parameters is a very old problem, and a number of techniques have been developed for dealing with it. The classical technique of estimation -- the method of least squares due to Gauss and others -- was developed by the early nineteenth century. The motivation for least squares, as described by Huber, is interesting:

> "The original motivation for this method (due to Gauss) is somewhat circular: least squares estimates are optimal if the errors are independent identically distributed normal; on the other hand, Gauss assumed a normal error law because then the sample mean, which 'is generally accepted as a good estimate,' turns out to be optimal in the simplest special case...." (p. 799 of [15]; also see p. 1042 of [14]).

The linchpin in this justification is, of course, the faith placed in the arithmetic mean. It is interesting to contrast the historical

dominance of least squares theory with the number of alternative techniques suggested over those 150 years which never gained prominence and which have only recently been resurrected. (For a history of these procedures and others, see [28].) Already in 1818 Laplace proposed, for estimating regression through the origin, minimizing the sum of absolute residuals rather than the sum of squared residuals. In the 1840s, in a paper criticizing Gauss' original justification for least squares, Ellis proposed essentially what are now called M estimators.

In the last several decades, the prominent position of least squares and the accompanying classical normal theory has been vigorously questioned. One of the most recent assessments of the possibility of poor performance of classical least squares theory comes from the Princeton robustness study (see [3]), which evaluated dozens of different estimators in the simple location problem. Their answer to the question "Which was the worst estimator in the study?" was: "If there is any clear candidate for such an overall statement, it is the arithmetic mean...." (p. 239 of [3]).

1.2.2 As was noted earlier, for the problem of estimating parameters in the regression models (1.1) and (1.3), a number of alternatives to the method of least squares have been developed. One such estimator for simple linear regression was originally proposed by Theil in 1950 [30] and later elaborated by Sen [27], who derived its asymptotic properties. Based on a rank test, the Sen-Theil estimator provides

an estimate of only the slope parameter β. In the simplest case where all the c_i are different, the estimate of β is simply the median of the $\binom{N}{2}$ slopes $(Y_j - Y_i)/(c_j - c_i)$ joining pairs of points.

There are also the estimators constructed to be analogues to the L, M, and R estimators for location. The most intensely studied class of analogues is the M-type estimators -- papers on this class include ones by Relles [26], Huber [15], Andrews [2], Bickel [6], and Yohai [31]. In this case (1.5) generalizes to: the estimator $\hat{\underset{\sim}{\beta}}{}^*$ of $\underset{\sim}{\beta}{}^*$ is that vector which causes

$$(1.7) \qquad \sum_{i=1}^{N} \rho(Y_i - \underset{\sim}{c}{}_i^{*T} \hat{\underset{\sim}{\beta}}{}^*) = \text{minimum.}$$

Obviously one member of this class is the least squares estimator: take $\rho(x) = x^2$. A more robust proposal for ρ given by Huber in [15] is

$$(1.8) \qquad \rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| < c \\ c|x| - \frac{1}{2}c^2 & |x| \geq c \end{cases},$$

where c is a constant. An entire family of M estimators is defined by

$$(1.9) \qquad \rho(x) = |x|^a \qquad \text{for } 1 \leq a \leq 2.$$

This family contains both the least squares estimator (a=2) and the estimator proposed by Laplace (a=1). Forsythe [9] and others have studied members of this family and the family as a whole.

In the published literature there appears to be only one analogue of an L estimator for regression, that of Bickel [5]. The analogy with the L estimates for location is more tenuous, however, than is the straightforward analogy obtaining for M estimates. Unlike the L estimate for location, Bickel's estimator requires a preliminary "reasonable" estimate of $\underset{\sim}{\beta}^*$ in order to form a type of one step improvement. On the other hand the asymptotic results Bickel derives for his analogue are identical (except of course for the dependence on the design matrix (c_{ij})) to those which hold in the location case. An interesting special case of Bickel's estimator is the analogue to the trimmed mean: the residuals derived from fitting the model with the preliminary estimate are ordered and a "position index" is associated with each; one trims those observations leading to residuals with extreme position indices and then forms the standard least squares estimate from the remaining observations (cf. pp. 599-601 of [5] for details).

Several estimators have been developed for the regression problem based on rank tests. We have already mentioned Sen-Theil; in addition there are R-type estimators proposed by Adichie [1], Koul [22], Kraft and van Eeden [23], Jurečková [21], and Jaeckel [18]. The oldest of these proposals is that of Adichie, who is concerned with estimation in the simple linear regression model. His estimators $\hat{\alpha}$ and $\hat{\beta}$ are constructed in a manner very similar to that of the Hodges-Lehmann estimator for shift: one forms rank test statistics T_1 and T_2 for

testing hypotheses on α and β respectively, and then by "inverting" these tests one derives the values for $\hat{\alpha}$ and $\hat{\beta}$ (cf. pp. 895-896 of [1] for details). There is an error in Adichie's paper which will assume some importance for us later: for Adichie's methods of proof to work, a necessary condition on the score function $\psi_0(u)$ used to generate the rank statistic T_2 is missing in the assumptions for the asymptotic normality of $\hat{\beta}$ (pp. 894-895 and p. 898). Specifically he needs to assume that ψ_0 be monotone increasing to insure that his conditions (A) and (B) (p. 895) obtain.

For the general regression problem, Jurečková [21] considered generalizations of the rank tests used by Adichie in order to define her estimates of the regression parameters. Her basic approach is to first estimate the vector $\underset{\sim}{\beta}$, and then to estimate α using a location estimate on the resulting residuals. For estimating $\underset{\sim}{\beta}$ she defines, for the sample $(Y_1,\ldots,Y_N)$ and a score generating function $J(u)$, the rank statistics

$$(1.10) \qquad S_{Nj}(\underset{\sim}{b}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} (c_{ij}-\bar{c}_j)\, a_N(R_i^{\underset{\sim}{b}}) \qquad j=1,\ldots,q,$$

with $\bar{c}_j = N^{-1} \sum_{i=1}^{N} c_{ij}$, $\underset{\sim}{b}^T = (b_1,\ldots,b_q)$, $a_N(k) = J\left(\frac{k}{N+1}\right)$, and $(R_1^{\underset{\sim}{b}},\ldots,R_N^{\underset{\sim}{b}})^T$ is the vector of ranks corresponding to the variables $Y_i - \underset{\sim}{c}_i^T \underset{\sim}{b}$, $i=1,\ldots,N$. (We note that in the case of simple linear regression, $q=1$, Jurečková's rank statistic S_N corresponds to Adichie's rank statistic T_2.) Jurečková's estimate $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ is then any value of

$\underset{\sim}{b}$ which minimizes

$$\sum_{j=1}^{q} |S_{Nj}(\underset{\sim}{b})| \tag{1.11}$$

For the case q=1 this definition leads to an estimate of β slightly different from Adichie's. Another estimator, closely related to Jurečková's but of a different motivation, is discussed next.

1.3 Jaeckel's estimator

1.3.1 Jaeckel's method of estimation is concerned exclusively with estimating the vector of regression parameters $\underset{\sim}{\beta}$ in the general regression model -- the technique does not directly admit an estimate of the intercept α; however, as in Jurečková's case, after the estimate $\tilde{\underset{\sim}{\beta}}$ is computed, α may be estimated by applying a robust estimate of location to the residuals $\{Y_i - \underset{\sim}{c}_i^T \tilde{\underset{\sim}{\beta}}\}_{i=1}^{N}$. The starting point for defining Jaeckel's estimate is a measure of the dispersion of a set of numbers; given this measure of dispersion, one constructs the residuals using the different possible values of $\underset{\sim}{\beta}$ and chooses as one's estimate the vector minimizing the dispersion of the residuals. This procedure has many elements in common with the M-estimators discussed earlier. Indeed, if for example one were to define the dispersion of the vector $\underset{\sim}{z}^T = (z_1, \ldots, z_N)$ as $\sum_{i=1}^{N} z_i^2$, then the procedure outlined above simply leads to the least squares estimates. What differentiates Jaeckel's procedure from the M-estimators, however, is his definition of the dispersion. Jaeckel defines the dispersion function $D_N(\underset{\sim}{z})$ as

$$D_N(\underset{\sim}{z}) = \sum_{k=1}^{N} a_N(k)\, z_{(k)} \quad , \tag{1.12}$$

where $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(N)}$ and the $\{a_N(k)\}$ are a set of <u>scores</u> satisfying

$$\sum_{k=1}^{N} a_N(k) = 0.$$

Thus for a vector of observations $\underset{\sim}{Y}^T = (Y_1, \ldots, Y_N)$ and vector $\underset{\sim}{b}$ and design matrix $C = (c_{ij})$, the dispersion is

$$D_N(\underset{\sim}{Y} - C\underset{\sim}{b}) = \sum_{k=1}^{N} a_N(k)\, [Y_{i(k)} - c_{i(k)}^T \underset{\sim}{b}], \tag{1.13}$$

where the bracketed quantity is notation for the k^{th} ordered residual. The estimate of $\underset{\sim}{\beta}$ is then any value of $\underset{\sim}{b}$ which minimizes $D_N(\underset{\sim}{Y} - C\underset{\sim}{b})$.

(Note on notation: In the remainder of the paper, we will distinguish the true value of the parameter $\underset{\sim}{\beta}$ (or β) by denoting it as $\underset{\sim}{\beta}_0$ (or β_0); similarly α_0 will denote the true value of α -- so model (1.3) reads

$$Y_i = \alpha_0 + \beta_0 c_i + e_i \ , \tag{1.3}$$

and so forth. Also, when no confusion will arise, we will shorten $D_N(\underset{\sim}{Y} - C\underset{\sim}{\beta})$ to $D_N(\underset{\sim}{\beta})$ to emphasize the dispersion as a function of $\underset{\sim}{\beta}$.)

<u>1.3.2</u> There are several motivations for Jaeckel's estimator. The first is that incorporated into the title of Jaeckel's paper: the idea

of minimizing the "dispersion" of the residuals. This, as we noted, is the same idea as involved in least squares and all M estimators: all Jaeckel has done is to choose another, but still intuitively appealing, formula for measuring how far a hypothesized line (or hyperplane in general) is from a set of observed values $(c_1,Y_1),\ldots,(c_N,Y_N)$ (or C and $\underset{\sim}{Y}$ in general).

The second motivation is not so apparent. But as Jaeckel proved, the asymptotic behavior of his estimator and that of Jurečková's is <u>identical</u>. This coincidence, it turns out, derives from the fact that, when it exists, the derivative of $D_N(\underset{\sim}{\beta})$ with respect to the $\beta_j{}^s$ is (except for a constant) equal to the set of Jurečková's rank statistics $\{S_{Nj}\}$. And so the good results for the asymptotic behavior of Jurečková's estimator carry over to Jaeckel's.

A third possible motivation for Jaeckel's estimator, which he did not mention, is related to the asymptotic behaviors (as functions of $\underset{\sim}{\beta}$) of the dispersion and of the log likelihood. We briefly consider a heuristic derivation of this relationship. We note that Jaeckel's results and proofs are unrelated to these heuristics and do not provide any indication of under what conditions they can be formalized. Suppose that the errors $\{e_i\}$ are iid with cdf F and density F'=f. We assume we use the scores generated by the function $\phi_f(u) = \frac{-f'}{f}\left(F^{-1}(u)\right)$ (i.e. $a_N(k) = \phi_f\left(\frac{k}{N+1}\right)$, k=1,...,N); for errors with cdf F this turns to be the optimal choice under certain restrictions. For simplicity we take q=1 (simple linear regression); and since the dispersion can

only be used to estimate β_0 we take $\alpha_0=0$. Then letting $L_N(\beta;\underset{\sim}{Y}) = L_N(\beta)$ be the likelihood based on N observations, we have

$$(1.14) \qquad \log L_N(\beta) = \sum_{i=1}^{N} \log f(e_i^\beta) = \sum_{i=1}^{N} \log f(e_{(i)}^\beta) \quad ,$$

where $e_i^\beta = Y_i - \beta c_i$ and $e_{(i)}^\beta$ is the i^{th} ordered value among $\{Y_k - \beta c_k\}_{k=1}^N$. We consider the behavior of $\log L_N(\beta)$ in the vicinity of the true value β_0. For $\beta \simeq \beta_0$ and N large, $e_{(i)}^\beta \simeq e_{(i)} \simeq F^{-1}(i/(N+1))$; expanding $\log f(e_{(i)}^\beta)$ about $F^{-1}(i/(N+1))$ in a Taylor series we get

$$(1.15) \qquad \log f(e_{(i)}^\beta) = \log f(F^{-1}(i/(N+1))) + \left[e_{(i)}^\beta - F^{-1}(i/(N+1))\right] \frac{f'}{f} \left[F^{-1}(i/(N+1))\right] + \text{higher order terms} \quad ;$$

thus on summing we get

$$\log L_N(\beta) = k_N - D_N(\beta) + R_N(\beta) \quad ,$$

where k_N is a constant independent of β and $R_N(\beta)$ is the sum of the higher order terms. Thus we see that, at least locally at β_0, $-D_N(\beta)$ and the log likelihood have similar asymptotic behavior if we use the correct scores in defining the dispersion. If the global behavior of $D_N(\beta)$ is reasonable, then we might expect similar asymptotic behavior for the maximum likelihood estimator and Jaeckel's estimator.

A fourth motivation for Jaeckel's estimate relates to its computational feasibility. First, unlike several other recent proposals (cf. [5], [23]), Jaeckel's estimator does not require a preliminary estimate of $\underset{\sim}{\beta}$. Second, Jaeckel's estimator seems to be easier to compute than Jurečková's proposal [21]. As Jaeckel points out, if the scores $a_N(k)$ are monotone increasing, then the dispersion is a continuous, <u>convex</u> function of $\underset{\sim}{\beta}$; also it is straightforward to compute the derivatives of the dispersion, which exist almost everywhere. Thus one can apply iterative methods for searching for the minimum; Jaeckel mentions the possibility of using the method of steepest descent.

One of the strengths of Jaeckel's estimator is its asymptotic performance as compared to that of other proposed estimators in use. As an example Jaeckel considered the simple linear regression case and compared the performances of the well-known Sen-Theil estimator and his estimator. Since the distribution of the errors is generally unknown, he chose for his scores $a_N(k)$ those optimal for the logistic distribution (Wilcoxon scores): $a_N(k) = \frac{k}{N+1} - \frac{1}{2}$. For this choice his estimate becomes a "weighted median" of the pairwise slopes $\left\{\frac{Y_j - Y_i}{c_j - c_i}\right\}$. With this set-up the asymptotic variance of the Sen-Theil estimator is always greater than (or equal to) that of Jaeckel's estimator.

<u>1.3.3</u> There are several areas of weakness in the results for the

estimators of Jaeckel, Jurečková, and Adichie. The first is the requirement in all three theories that the scores be generated by a function $\phi(u)$ which is monotone increasing. This is an important restriction since for many choices of F, the distribution function of the errors, the "optimal" choice of scores -- $\phi_f(u) = \frac{-f'}{f}(F^{-1}(u))$ -- is not monotone increasing. If F is the Cauchy distribution, then $\phi_f(u) = \sin(2\pi u - \pi)$, which is not monotone for $u \varepsilon [0,1]$. Indeed, an easy calculation shows that if F is a t distribution with n degrees of freedom the corresponding $\phi_f(u)$ is not monotone for any choice of n. Also numerical results indicate that if F is one of a variety of contaminated normal distributions, then ϕ_f is not monotone either. It turns out that $\phi_f(u)$ is monotone if and only if $F'=f$ is a so-called "strongly unimodal" density (cf. [10]). Obviously for applications one would like to be able to use non-monotone scores and be assured of the asymptotic performance of the resulting estimator.

The second deficiency arises since, in practice, one seldom knows the distribution function F. In this case one may, of course, choose an omnibus score function ϕ, such as the Wilcoxon score function mentioned earlier; or if one has some idea of the shape of F, one can try to choose ϕ which works reasonably well (although not optimally) for all these feasible shapes. An alternative to choosing one specific score function is to choose the score function (possibly from a given family of choices) on the basis of the sample -- that is to make the estimator adaptive. In the location problem using L estimators,

adapting has been used quite successfully: not only has it provided estimators which are (nearly) asymptotically optimal over a large nonparametric family of error distributions, but the estimators also perform well in samples of modest size. (For details on adapting L estimators, see Johns [19].) It would be very useful if adaptive estimators with similar properties could be found in the regression problem.

1.4 Outline of Results

The main results of this paper fall into three categories. The first category contains results related to the asymptotic consistency of Jaeckel's estimator when non-monotone score functions are used. In Section 2.2 conditions are stated for the problem of simple linear regression under which consistency obtains together with a proof of the result, which is similar in spirit to the classic proof by Wald of the consistency of maximum likelihood estimates. Section 2.3 contains a counterexample to the consistency of Jaeckel's estimator based on non-monotone scores if certain of the conditions on the error distribution, invoked in Section 2.2, are not met. It should be noted that these conditions are not necessary if one uses monotone scores.

In proving the results of Sections 2.2 and 2.3, we utilize results due to Stigler [29] on the behavior of linear combinations of order statistics. However some of the results stated in his paper are incorrect and there are errors and gaps in some of his proofs.

The second category of results, contained in Sections 2.4 through 2.7, corrects these deficiencies so the results can be used in the proofs concerning consistency.

The third category of results, comprising Chapter 3, addresses the difficulty of not knowing the true distribution of the errors in the general regression model. An adaptive estimator is proposed based on a family of Jaeckel-type estimators (with monotone scores), and results are proved concerning its asymptotic behavior. These results show that, asymptotically at least, one loses very little in not knowing the error distribution (if it is strongly unimodal) if one uses this adaptive estimator.

CHAPTER 2. CONSISTENCY OF JAECKEL'S ESTIMATOR FOR NONMONOTONE SCORES

2.1 Model and Assumptions

Throughout this chapter we will be concerned with simple linear regression through the origin:

$$Y_i = \beta_0 c_i + e_i \quad , \tag{2.1}$$

where $Y_1,\ldots,Y_N$ are observations on the dependent variable, $c_1,\ldots,c_N$ are regression constants, β_0 is the unknown slope parameter to be estimated, and $e_1,\ldots,e_N$ are iid random variables with distribution function F. We then define Jaeckel's dispersion function for the score-generating function $J(u)$ ($J: [0,1] \to R$) by

$$D_N(\beta) = \sum_{k=1}^{N} a_N(k)\, e_{(k)}^{\beta} \quad , \tag{2.2}$$

where $a_N(k) = J(\frac{k}{N+1})$ and $e_{(k)}^{\beta}$ is the k^{th} ordered value among the residuals $\{Y_i - \beta c_i : i=1,\ldots,N\}$, as defined earlier on p.12. Note that in the definition of D_N, the dependence on the $\{Y_i\}$ and $\{c_i\}$ has been suppressed. Jaeckel's estimate of β_0 is denoted by $\tilde{\beta}_N$ and is any value of β in the parameter space B^o satisfying

$$D_N(\tilde{\beta}_N) \leqslant D_N(\beta) \quad \text{for all } \beta \in B^o \quad . \tag{2.3}$$

Assumptions

We will now summarize the usual assumptions we will invoke in the course of this chapter.

F1: F is unimodal and symmetric about 0.

F2: F has density f with $f \leq B_f$, B_f a finite constant. Also F has finite Fisher's information $I(f) = \int [f'/f]^2 f$.

F3: (Tail condition) For some $\varepsilon_1 > 0$, $\lim_{x \to \infty} x^{\varepsilon_1} [1 - 2F(x)] = 0$.

B1: $B^o \subseteq R$ is a compact set; without loss of generality take $B^o = [-B, B]$.

C1: Define $H_N(x) = N^{-1} \times \#\{c_i \leq x: \ i = 1, \ldots, N\}$, where $\#A$ denotes the cardinality of a set A. Thus H_N is the "sample distribution function" of the $\{c_i: \ i = 1, \ldots, N\}$. We assume $|c_i| \leq B_c$ for all i, so H_N concentrates all its mass on $[-B_c, B_c]$, and that $H_N \xrightarrow{w}$ some H, a distribution function. We assume the variance of H, $\mathrm{var}(H)$, is strictly positive.

J1: $0 \leq J(\tfrac{1}{2} + u) = -J(\tfrac{1}{2} - u)$ for $u \in [0, \tfrac{1}{2})$, with $J(\tfrac{1}{2} + u) > 0$ on some interval.

J2: $|J(u)| \leq B_J < \infty$ for $u \in [0,1]$. Also, J satisfies a Hölder condition with $\gamma > \tfrac{1}{2}$ (i.e. for $u, v \in [0,1]$,

$$|J(u) - J(v)| \leq \text{constant} \cdot |u - v|^{\gamma} \ ),$$

except possibly at a finite set of points $t_1, \ldots, t_p$. Lastly J trims: for some $\alpha \in (0, \tfrac{1}{2})$, $J(u) = 0$ if $u \leq \alpha$ or $u \geq 1-\alpha$.

Jaeckel's Theorem

For the sake of reference, we state Jaeckel's asymptotic normality result:

Theorem Let F have finite Fisher's information and suppose that J(u) is non-constant, non-decreasing, and square integrable on (0,1) such that $\sum_{k=1}^{N} a_N(k) = 0$. Then under some technical assumptions on the $\{c_i\}$ (cf. p. 1328 of [21]),

$$N^{\frac{1}{2}} (\tilde{\beta}_N - \beta_0) \xrightarrow{D} N(0,V), \text{ where } \tilde{\beta}_N \text{ is Jaeckel's}$$

estimator and

$$V = \frac{\int_0^1 (J(u)-\bar{J})^2 du}{\left[\int_0^1 J(u)\phi_f(u)\, du\right]^2} \cdot \Sigma^{-1}, \text{ with}$$

$\bar{J} = \int_0^1 J(u)du$, $\phi_f(u) = -f'/f(F^{-1}(u))$, and $\Sigma = [\sigma_{lj}]$ with

$$\sigma_{lj} = \lim_{N\to\infty} N^{-1} \cdot \sum_{k=1}^{N} (c_{kl} - \bar{c}_l)(c_{kj} - \bar{c}_j) \text{ and } \bar{c}_j = N^{-1} \cdot \sum_{k=1}^{N} c_{kj}.$$

If q=1 (simple linear regression), Σ is a scalar equal to var(H).§

2.2 Consistency Proof

In this section we prove the consistency of Jaeckel's estimator $\tilde{\beta}_N$ for the true slope parameter β_0. Because of the invariance of Jaeckel's estimator (cf. p. 1452 of [18]), we will assume without loss of generality that $\beta_0 = 0$ in the remainder of the paper. The method of proof will be to use Theorem 7* of Section 2.6 to derive asymptotic

results about the behavior of the dispersion $D_N(\beta)$ and then use the compactness of the set B^o of possible values of the slope parameter.

For ease of reference we now state a version of Theorem 7* which we will use in this section ($\sigma^2(Z)$ is an alternate notation for the variance of a random variable Z):

<u>Theorem 7*</u> (special version) Let $X_{1N},\dots, X_{NN}$ be independent random variables with cdf^s $F_{1N},\dots, F_{NN}$ respectively; suppose that for some cdf $G(y)$, for which there is an $\varepsilon_1 > 0$ such that $\lim_{x\to\infty} x^{\varepsilon_1}\left[1-G(x)-G(-x)\right] = 0$, there is a finite constant M such that if $y \leq -M$ then $F_{kN}(y) \leq G(y)$ and if $y \geq M$ then $F_{kN}(y) \geq G(y)$ (for all k,N). Assume that for a.e. x,y (with respect to Lebesgue measure) the following limits exist:

$$\lim_{N\to\infty} \bar{F}_N(x) \equiv \bar{F}(x) \ , \text{ where } \bar{F}_N(x) = N^{-1} \cdot \sum_{k=1}^{N} F_{kN}(x)$$

and $$\lim_{N\to\infty} N^{-1} \cdot \sum_{k=1}^{N} \{F_{kN}(\min(x,y)) - F_{kN}(x)\, F_{kN}(y)\} \equiv K(x,y).$$

Also assume that $\bar{F}_N^{-1}$ is absolutely continuous with respect to Lebesgue measure for each N. Let J be a score function satisfying Assumption J2 and define $S_N = N^{-1} \cdot \sum_{i=1}^{N} J(i/(N+1))\, X_{(i)}$, where $X_{(i)}$ is the i^{th} ordered value among $X_{1N},\dots, X_{NN}$. Then

(i) $\lim_{N\to\infty} N\sigma^2(S_N) = \sigma^2(J,\bar{F},K)$ (given below);

(ii) if $\sigma^2(J,\bar{F},K) > 0$, then

$$\frac{S_N - E(S_N)}{\sigma(S_N)} \xrightarrow{D} N(0,1) \quad \text{as } N\to\infty;$$

(iii) $N^{\frac{1}{2}}(E(S_N) - \mu(J,\bar{F}_N)) \to 0$ as $N\to\infty$ where

$$\mu(J,\bar{F}_N) = \int_0^1 J(u)\, \bar{F}_N^{-1}(u)\, du \quad \text{and}$$

$$\sigma^2(J,\bar{F},K) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\bar{F}(x))\, J(\bar{F}(y))\, K(x,y)\, dx\, dy. \S$$

We now consider the application of this result. Let the value of β be fixed. In order to use Theorem 7* we define the function

$$(2.4) \qquad G(y) = \begin{cases} F(y-BB_c) & y \geq BB_c \\ F(y+BB_c) & y < -BB_c \\ \text{arbitrarily in } [-BB_c, BB_c) \text{ so that} & \\ \quad G(y) \text{ is a distribtuion function.} & \end{cases}$$

Then

$$\lim_{x\to\infty} x^{\varepsilon_1}\left[1-G(x)+G(-x)\right] =$$

$$\lim_{x\to\infty} \{(x-BB_c)^{\varepsilon_1}\left[1-2F(x-BB_c)\right]\} \cdot \lim_{x\to\infty}\left[\frac{x}{x-BB_c}\right]^{\varepsilon_1} =$$

$0\cdot 1 = 0$ by Assumption F3, where ε_1 is as in that assumption.

Defining the resuidual $e_i^\beta = Y_i - \beta c_i$ and letting F_{iN} be the cdf of e_i^β, we note that the cdf^s $\{F_{iN}\}_{i=1}^N$ and G satisfy the required relationship in the assumptions of Theorem 7*, with the M of Theorem 7* being BB_c. Also, for all x,

$$(2.5) \qquad \bar{F}_{\beta,N}(x) \equiv N^{-1} \cdot \sum_{k=1}^{N} F_{kN}(x) = \int F(x+\beta c)\, dH_N(c) \to \bar{F}_\beta(x)$$

as $N\to\infty$, where $\bar{F}_\beta(x) \equiv \int F(x+\beta c)\, dH(c)$, by Assumption C1; similarly,

for all x and y,

$$N^{-1} \cdot \sum_{k=1}^{N} \left[F_{kN}(\min(x,y)) - F_{kN}(x)F_{kN}(y)\right] \to K_\beta(x,y) \tag{2.6}$$

as $N\to\infty$, where $K_\beta(x,y) \equiv \int\left[F(\min(x,y)+\beta c) - F(x+\beta c)F(y+\beta c)\right] dH(c)$.

Since F(x) is everywhere continuous and strictly increasing (for x^s such that $F(x)\varepsilon(0,1)$) by Assumptions F1 and F2, for each fixed β, $\bar{F}_{\beta,N}(x)$ and $\bar{F}_\beta(x)$ are also everywhere continuous and strictly increasing, implying $\bar{F}_{\beta,N}^{-1}$ and $\bar{F}_\beta^{-1}$ are absolutely continuous with respect to Lebesgue measure. Thus J is uniformly continuous a.e. $\bar{F}_\beta^{-1}$ and satisfies the Hölder condition a.e. $\bar{F}_{\beta,N}^{-1}$ for all N=1,2,... by Assumption J2. Thus Theorem 7* is applicable. Letting

$$S_N = N^{-1} \cdot \sum_{i=1}^{N} J(i/(N+1))\, e_{(i)}^{\beta}\ , \tag{2.7}$$

Theorem 7* implies (with $\sigma^2(W)$ denoting the variance of the r.v. W)

$$N\sigma^2(S_N) \longrightarrow \sigma^2(J,\bar{F}_\beta,K) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[J(\bar{F}_\beta(x))J(\bar{F}_\beta(y))K_\beta(x,y)\right]dxdy\ . \tag{2.8}$$

For simplicity we usually denote $\sigma^2(J,\bar{F}_\beta,K)$ by σ_β^2 to emphasize its dependence on β. We can assume $\sigma_\beta^2>0$, since otherwise equation (2.11) below follows immediately. Then by Theorem 7* we can conclude, since $D_N(\beta) = \dfrac{S_N}{N}$, that

$$(2.9) \qquad \text{(i)} \quad \left\{ \frac{\frac{D_N(\beta)}{N} - E\left(\frac{D_N(\beta)}{N}\right)}{\sigma\left(\frac{D_N(\beta)}{N}\right)} \right\} \xrightarrow{D} N(0,1) \quad \text{and}$$

$$\text{(ii)} \quad N^{\frac{1}{2}} \left\{ E\left(\frac{D_N(\beta)}{N}\right) - \mu(J,\bar{F}_{\beta,N}) \right\} \longrightarrow 0 \quad \text{as}$$

$N \to \infty$, where $\mu(J,\bar{F}_{\beta,N}) \equiv \int_0^1 J(u)\, \bar{F}^{-1}_{\beta,N}(u)du$ and E denotes expectation. We usually denote $\mu(J,\bar{F}_{\beta,N})$ by $\mu_N(\beta)$ and $\mu(J,\bar{F}_\beta)$ by $\mu(\beta)$. Combining the results of (2.8) and (2.9):

$$(2.10) \qquad \left\{ \sqrt{N} \left[\frac{\frac{D_N(\beta)}{N} - \mu_N(\beta)}{\sigma_\beta} \right] \right\} \xrightarrow{D} N(0,1) \quad .$$

By a standard result of Mann-Wald theory (cf. [24] for the result and notation), if a sequence of random variables X_n converges in distribution to a random variable X, then $X_n = 0_p(1)$. Thus by (2.10)

$$\sqrt{N} \left\{ \frac{\frac{D_N(\beta)}{N} - \mu_N(\beta)}{\sigma_\beta} \right\} = 0_p(1) \quad ,$$

which implies $D_N(\beta) = N\mu_N(\beta) + 0_p(\sqrt{N})$ and $D_N(0) = N\mu_N(0) + 0_p(\sqrt{N})$. Combining we get

$$(2.11) \qquad D_N(\beta) = D_N(0) + N(\mu_N(\beta) - \mu_N(0)) + 0_p(\sqrt{N}) \quad \text{for each } \beta.$$

To proceed farther in proving consistency, we need the behavior of $\mu_N(\beta) - \mu_N(0)$. Specifically we show that for any fixed $\beta \neq 0$ there is $\eta > 0$ such that $\mu_N(\beta) - \mu_N(0) > \eta$ for all N sufficiently large. We do this in several steps.

Lemma 2.1 $\quad \mu(\beta) = \mu(-\beta)$.§

Proof
$$\mu(\beta) = \int_0^1 J(u)\, \bar{F}_\beta^{-1}(u)\, du$$
$$= \int_0^{\frac{1}{2}} J(\tfrac{1}{2}+u) \left[\bar{F}_\beta^{-1}(\tfrac{1}{2}+u) - \bar{F}_\beta^{-1}(\tfrac{1}{2}-u)\right] du$$

since $J(\frac{1}{2}-u) = -J(\frac{1}{2}+u)$ by Assumption J1. Hence to show $\mu(\beta) = \mu(-\beta)$ it suffices to prove

$$\text{(2.12)} \qquad \bar{F}_\beta^{-1}(\tfrac{1}{2}+u) - \bar{F}_\beta^{-1}(\tfrac{1}{2}-u) = \bar{F}_{-\beta}^{-1}(\tfrac{1}{2}+u) - \bar{F}_{-\beta}^{-1}(\tfrac{1}{2}-u) \quad \text{for all } u \in [0, \tfrac{1}{2}).$$

As before $\bar{F}_\beta$ is strictly increasing, implying $\bar{F}_\beta$ has a unique inverse. Let $w = \bar{F}_\beta^{-1}(\frac{1}{2}+u)$, so $\bar{F}_\beta(w) = \frac{1}{2}+u$. Then

$$\tfrac{1}{2}+u = \int_{-B_c}^{B_c} F(w+\beta c)\, dH(c)$$

$$= \int_{-B_c}^{B_c} \left[1 - F(-w-\beta c)\right] dH(c) \quad \text{(since f is symmetric)}$$

$$= 1 - \int_{-B_c}^{B_c} F(-w-\beta c)\, dH(c) \quad ; \quad \text{thus}$$

$$\tfrac{1}{2}-u = \int_{-B_c}^{B_c} F(-w-\beta c)\, dH(c)$$

$$= \bar{F}_{-\beta}(-w) \quad , \quad \text{implying}$$

$$\bar{F}_{-\beta}^{-1}(\tfrac{1}{2}-u) = -w, \qquad \text{so}$$

$$\bar{F}_{\beta}^{-1}(\tfrac{1}{2}+u) = -\bar{F}_{-\beta}^{-1}(\tfrac{1}{2}-u) \quad . \tag{2.13}$$

This argument also shows $\bar{F}_{\beta}^{-1}(\tfrac{1}{2}-u) = -\bar{F}_{-\beta}^{-1}(\tfrac{1}{2}+u)$, which combined with (2.13) yields (2.12). Thus $\mu(\beta) = \mu(-\beta)$. §

Because of this result we will now restrict our attention to $\beta>0$. Indeed, throughout the rest of this chapter we will restrict attention to $\beta>0$, since the results for $\beta<0$ follow in an analogous fashion.

We define:

$$C(u) \equiv \bar{F}_{\beta}^{-1}(\tfrac{1}{2}-u) - \bar{F}_{\beta}^{-1}(\tfrac{1}{2}+u) + 2F^{-1}(\tfrac{1}{2}+u). \tag{2.14}$$

Then we have

Lemma 2.2 $\quad \mu(\beta) - \mu(0) = -\int_0^{\frac{1}{2}} J(u+\tfrac{1}{2})\, C(u)\, du.$ §

Proof $\quad$ Note that $\bar{F}_0(u) = \int_{-B_c}^{B_c} F(u)\, dH(c) = F(u)$, so $\bar{F}_0^{-1}(u) = F^{-1}(u)$. Thus $\mu(\beta) - \mu(0) = \int_0^1 J(u)\left[\bar{F}_{\beta}^{-1}(u) - F^{-1}(u)\right]du$; the result easily follows on recalling $J(\tfrac{1}{2}-u) = -J(\tfrac{1}{2}+u)$. §

Lemma 2.3 $\quad$ If $\beta\neq 0$ then $\mu(\beta)>\mu(0)$. §

Proof $\quad$ By Assumption J1 $J(\tfrac{1}{2}+u)\geqslant 0$ for $u\in[0,\tfrac{1}{2}]$ with strict inequality on some interval; so to prove the result it suffices to show $C(0) = 0$

and $C(u)<0$ for $u \varepsilon (0,\frac{1}{2}]$. Now $C(0) = \bar{F}_\beta^{-1}(\frac{1}{2}) - \bar{F}_\beta^{-1}(\frac{1}{2}) + 2F^{-1}(\frac{1}{2}) = 0$

by the symmetry of f. Let $u \varepsilon (0,\frac{1}{2}]$ and set $d = F^{-1}(\frac{1}{2}+u) > 0$. Suppose x is defined by $\bar{F}_\beta(x) = F(d)$, so $x = \bar{F}_\beta^{-1}(\frac{1}{2}+u)$. Then

$$\begin{aligned} C(u) &= \bar{F}_\beta^{-1}(\tfrac{1}{2}-u) - \bar{F}_\beta^{-1}(\tfrac{1}{2}+u) + 2F^{-1}(\tfrac{1}{2}+u) \\ &= \bar{F}_\beta^{-1}(\tfrac{1}{2}-u) - x + 2d. \end{aligned}$$

If we denote $y = \bar{F}_\beta^{-1}(\frac{1}{2}-u)$ and $y_0 = x-2d$, to show $C(u)<0$ it will suffice

to show $y<y_0$, or equivalently

$$\bar{F}_\beta(y_0) = \bar{F}_\beta(x-2d) > \tfrac{1}{2}-u \quad , \tag{2.15}$$

since $\bar{F}_\beta$ is strictly increasing. To continue the proof we need the following

<u>Fact</u>: If $z \neq -d$, then $F(-d) - F(z) < F(d) - F(2d+z)$.

<u>Proof of Fact</u>: Suppose $z<-d$. Then (considering Figure 1)

$$\begin{aligned} F(-d) - F(z) &= \int_z^{-d} f(u)\, du \\ &= \int_d^{-z} f(u)\, du \end{aligned}$$

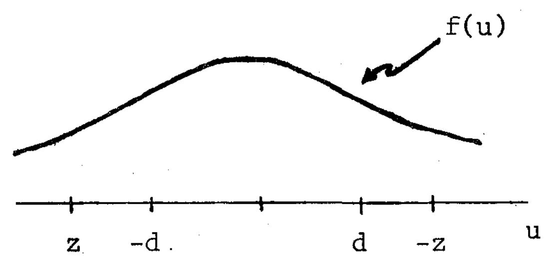

FIGURE 1

(by symmetry), so

$$F(-d) - F(z) = \int_0^{-z-d} f(v+d)\, dv. \tag{2.16}$$

Since f is unimodal and d>0, $f(d+v)<f(d-v)$ for all $v>0$, implying

$$\int_0^{-z-d} f(v+d)\,dv < \int_0^{-z-d} f(d-v)\,dv$$

$$= \int_{2d+z}^{d} f(u)\,du$$

$$= F(d) - F(2d+z) \;.$$

The same argument shows that if $z>-d$, then

$$F(z) - F(-d) > F(2d+z) - F(d), \quad \text{or equivalently}$$

$$F(-d) - F(z) < F(d) - F(2d+z) \;,$$

completing the proof of the Fact.

Now $$\bar{F}_\beta(y_0) = \int_{-B_c}^{B_c} F(y_0+\beta c)\,dH(c)$$

$$= \int_{y_0-\beta B_c}^{y_0+\beta B_c} F(z)\,dH((z-y_0)/\beta) \quad .$$

By the Fact, $F(z) > [1-2F(d)] + F(2d+z)$, so

$$\bar{F}_\beta(y_0) > \int_{y_0-\beta B_c}^{y_0+\beta B_c} [1-2F(d)]\,dH((z-y_0)/\beta)$$

$$+ \int_{y_0-\beta B_c}^{y_0+\beta B_c} F(2d+z)\,dH((z-y_0)/\beta) \quad .$$

The first integral is just $1-2F(d)$.

The second integral $$= \int_{-B_c}^{B_c} F(2d+y_0+\beta c)\,dH(c)$$

$$= \int_{-B_c}^{B_c} F(x+\beta c)\,dH(c)$$

$$= F(d).$$

Thus $\bar{F}_\beta(y_0) > 1-2F(d) + F(d) = \frac{1}{2}-u$, implying (2.15); thus Lemma 2.3 is proved. §

Lemma 2.4 Let $\beta \neq 0$; then there exists $\eta(\beta)>0$ and N^* such that $\mu_N(\beta) - \mu_N(0) > \eta(\beta)$ for all $N \geq N^*$. §

Proof We first note that $\mu_N(0) = \mu(0)$. Also, by Lemma 2.3 $\mu(\beta) - \mu(0)>0$. Hence it will suffice to show

$$\lim_{N\to\infty} \mu_N(\beta) = \mu(\beta), \qquad \text{i.e. that}$$

$$\int_0^1 J(u)\,\bar{F}^{-1}_{\beta,N}(u)\,du \longrightarrow \int_0^1 J(u)\,\bar{F}^{-1}_{\beta}(u)\,du \quad \text{as } N\to\infty. \tag{2.17}$$

Since $J(u) = 0$ for $u\leqslant\alpha$ and $u\geqslant 1-\alpha$, the dominated convergence theorem will imply (2.17) if we show:

(i) $\bar{F}^{-1}_{\beta,N}(u) \to \bar{F}^{-1}_{\beta}(u)$ for $u\varepsilon[\alpha,1-\alpha]$, and

(ii) there is $\nu(u)$ integrable on $[\alpha,1-\alpha]$ such that

$$|\bar{F}^{-1}_{\beta,N}(u)|\leqslant\nu(u) \text{ for } u\varepsilon[\alpha,1-\alpha] \text{ for all } N \text{ large.}$$

Proof of (i): By assumption C1, $\bar{F}_{\beta,N}(x) \to \bar{F}_\beta(x)$ for all x. To simplify notation let $g_N(x) = \bar{F}_{\beta,N}(x)$ and $g(x) = \bar{F}_\beta(x)$. Then we know $g_N(x) \longrightarrow g(x)$ for all x, and we wish to prove $g_N^{-1}(u) \to g^{-1}(u)$ for all $u\varepsilon[\alpha,1-\alpha]$. Let $\varepsilon>0$ be given and pick $u\varepsilon[\alpha,1-\alpha]$; denote $x=g^{-1}(u)$. We must show $|g_N^{-1}(u)-x|<\varepsilon$ for all N large. By the monotonicity of g_N it will suffice to show (for N large):

$$(2.18) \qquad \begin{cases} g_N(x-\varepsilon)<u \\ g_N(x+\varepsilon)>u \end{cases} .$$

But $g_N(x-\varepsilon) \to g(x-\varepsilon)<u$, and $g_N(x+\varepsilon) \to g(x+\varepsilon)>u$, so for all N large (2.18) obtains, implying (i).

Proof of (ii): For $u\varepsilon[\alpha,1-\alpha]$, $|g_N^{-1}(u)|\leq\max\{|g_N^{-1}(1-\alpha)|, |g_N^{-1}(\alpha)|\}$. But by (i) $g_N^{-1}(1-\alpha) \to g^{-1}(1-\alpha)$ and $g_N^{-1}(\alpha) \to g^{-1}(\alpha)$ as $N\to\infty$. Therefore set $\nu(u) = \max\{|g^{-1}(1-\alpha)| +1, |g^{-1}(\alpha)|+1\}$. Then for all N sufficiently large $|\bar{F}_{\beta,N}^{-1}(u)|\leq\nu(u)$. Trivially ν is integrable on $[\alpha,1-\alpha]$, implying (ii).

Thus (2.17) obtains, completing the proof of Lemma 2.4. §

Lemma 2.5 Whenever it exists, the derivative of the dispersion has the bound $|D_N'(\beta)| \leq NB_JB_c$. §

Proof By Jaeckel's Theorem 1 and the remarks of p. 1451 of [18], $D_N(\beta)$ is a non-negative, continuous, piecewise linear function of β (even in the case J(u) is not increasing for $u\varepsilon[\frac{1}{2},1]$). By p. 1455 of [18], where it exists

$$D_N' (\beta) = - \sum_{k=1}^{N} a_N(k)\, c^{(k)} = - \sum_{k=1}^{N} J(k/(N+1))\, c^{(k)},$$

where $c^{(k)}$ is the c value associated with the residual $e^{\beta}_{(k)}$. Hence

$$|D_N' (\beta)| \leq \sum_{k=1}^{N} |J(k/(N+1))|\, |c^{(k)}| < NB_JB_c \quad \text{by}$$

Assumptions C1 and J2. §

Theorem 2.1 (Consistency of Jaeckel's estimator)

Assume $\beta_0=0$ without loss of generality. Then under the assumptions of Section 2.1,

$$\tilde{\beta}_N \xrightarrow{P} 0. \ \S$$

Proof The idea of the proof is the following simple observation. For $\omega\varepsilon\Omega$, if $D_N(\beta)\ (\omega) > D_N(0)\ (\omega)$, then $\tilde{\beta}_N(\omega) \neq \beta$ since $\tilde{\beta}_N$ minimizes $D_N(\beta)$ over all $\beta\varepsilon B^0$. Let $\Delta,\varepsilon>0$ be arbitrary and choose β^* outside the interval $0_\Delta=(-\Delta,\Delta)$. Then by the piecewise linearity of D_N and Lemmas 2.4 and 2.5, there is an interval $I_{\beta^*}=(\beta^*-h_{\beta^*},\ \beta^*+h_{\beta^*})$ and N_{β^*} such that for all $N\geqslant N_{\beta^*}$

$$\sup_{\beta\varepsilon I_{\beta^*}} \left|D_N(\beta)-D_N(\beta^*)\right| < \tfrac{1}{2}N\cdot\left[\mu_N(\beta^*)-\mu(0)\right] \tag{2.19}$$

(simply choose $h_{\beta^*} < \eta(\beta^*)/(2B_J B_c)$). Consider the collection of open sets $\{I_{\beta^*}\colon \beta^*\varepsilon B^0\sim 0_\Delta\}$. This collection provides an open cover of $B^0\sim 0_\Delta$, which is compact by Assumption B1. Thus there is a finite subcover $I_{\beta_1},\ldots,I_{\beta_p}$ say. Recall that by (2.11),

$$D_N(\beta_i) = D_N(0) + N(\mu_N(\beta_i)-\mu(0)) + 0_p(\sqrt{N})$$

(where we note that the term $0_p(\sqrt{N})$ may depend on β_i). Thus for each β_i ($i=1,2,\ldots,p$), there exists N_i (for which (2.19) -- with $\beta^*=\beta_i$-- obtains for $N\geqslant N_i$) such that

$$P\{\left|D_N(\beta_i) - D_N(0) - N(\mu_N(\beta_i)-\mu(0))\right| \geqslant \tfrac{1}{2}N\cdot(\mu_N(\beta_i)-\mu(0))\} < \frac{\varepsilon}{p}$$

for all $N\geqslant N_i$. If we let $N^* = \max_{i=1,\ldots,p} \{N_i\}$, then for all $N\geqslant N^*$ the following p inequalities hold simultaneously with probability $> 1-\varepsilon$:

$$(2.20) \qquad |D_N(\beta_i) - D_N(0) - N(\mu_N(\beta_i)-\mu(0))| < \frac{N}{2}(\mu_N(\beta_i) - \mu(0)) \quad , \quad i = 1,2,\ldots,p.$$

Now any $\beta \varepsilon B^0 \sim O_\Delta$ satisfies $\beta \varepsilon I_{\beta_k}$ for some $k = 1,2,\ldots,p$; also for all $N \geqslant N^*$

$$|D_N(\beta) - D_N(\beta_k)| < \frac{N}{2}(\mu_N(\beta_k) - \mu(0))$$

by (2.19). Hence by (2.20) we obtain

$$P\{\inf_{\beta\varepsilon B^0 \sim O_\Delta} D_N(\beta) > D_N(0)\} > 1 - \varepsilon$$

for all $N \geqslant N^*$. Thus $P\{\tilde{\beta}_N \not\in (-\Delta,\Delta)\} < \varepsilon$ for all $N \geqslant N^*$, implying $\tilde{\beta}_N \xrightarrow{P} 0$ since ε and Δ were arbitrary. §

2.3 Counterexample

The aim of this section is to show that, in the case of a non-monotone score function $J(u)$, some sort of condition (not a regularity condition) needs to be invoked on the distribution of the errors in order that Jaeckel's estimator have the desired asymptotic properties. In the last section we proved the consistency of Jaeckel's estimator assuming f is unimodal; the counterexample that follows shows that this assumption cannot be dropped without invoking some other conditions. In the case of non-monotone scores, there are non-unimodal densities (otherwise well-behaved) for which Jaeckel's estimator is not consistent for the true slope parameter.

Theorem 2.2

There are a non-monotone score function $J(u)$ and a non-unimodal density f, with J satisfying the assumptions of Section 2.1 and f satisfying all but the unimodality assumption of Section 2.1, for which $\tilde{\beta}_N \overset{P}{\not\to} 0$. §

To carry out the proof we need

Lemma 2.6

There exist J and f as described in Theorem 2.2 and $\beta' \neq 0$ such that $\mu(\beta') < \mu(0)$. §

Proof Recall $\mu(\beta) - \mu(0) = -\int_0^{\frac{1}{2}} J(u+\frac{1}{2})\, C(u)\, du$. Also $C(0) = 0$. We will find β' and $\varepsilon>0$ such that $C(u) > 0$ for $u \in (0,\varepsilon)$. This will be sufficient to prove the lemma since we will then choose (cf. Figure 2) J such that $J(u+\frac{1}{2}) = 0$ for $u > \varepsilon$ and $J(u+\frac{1}{2}) > 0$ for $u \in (0,\varepsilon/2)$ say.

We'll assume H has a density h, symmetric about 0. Since

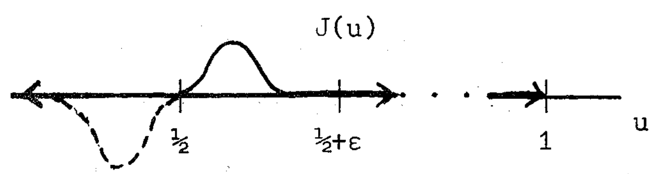

FIGURE 2

$$C(u) = \bar{F}_\beta^{-1}(\tfrac{1}{2} - u) - \bar{F}_\beta^{-1}(\tfrac{1}{2} + u) + 2F^{-1}(\tfrac{1}{2} + u),$$

$$C'(u) = \frac{2}{f(F^{-1}(\frac{1}{2} + u))} - \left\{ \int_{-B_c}^{B_c} f(\bar{F}_\beta^{-1}(\tfrac{1}{2}-u) + \beta c)\, h(c)\, dc \right\}^{-1} - \left\{ \int_{-B_c}^{B_c} f(\bar{F}_\beta^{-1}(\tfrac{1}{2}+u) + \beta c)\, h(c)\, dc \right\}^{-1} .$$

Therefore

$$C'(0) = 2\left[\frac{1}{f(0)} - \frac{1}{\int_{-B_c}^{B_c} f(\bar{F}_\beta^{-1}(\tfrac{1}{2}) + \beta c)\, h(c)\, dc}\right] .$$

If h is symmetric about 0 it is easy to show that $\bar{F}_\beta^{-1}(\tfrac{1}{2}) = 0$, so

$$C'(0) = 2\left[\frac{1}{f(0)} - \frac{1}{\int_{-B_c}^{B_c} f(\beta c)\, h(c)\, dc}\right] . \tag{2.21}$$

Now B_c is fixed. If $f(x) > f(0)$ in an open neighborhood of 0 (with 0 deleted), and $f(0) > 0$, then by taking β sufficiently close to 0 $(\beta \neq 0)$ we obtain

$$\int_{-B_c}^{B_c} f(\beta c)\, h(c) dc > f(0),$$

implying $C'(0) > 0$. Since C has a finite first derivative at $t = 0$, a theorem from calculus (cf. Chung [8], p. 156) implies

$$\begin{aligned} C(t) &= C(0) + C'(0)\cdot t + o(|t|) \quad \text{as } t \to 0 \\ &= tC'(0) + o(|t|) . \end{aligned} \tag{2.22}$$

Thus there is $\varepsilon > 0$ such that $0 < t < \varepsilon$ implies $C(t) > 0$. §

<u>Proof of Theorem 2.2</u> We consider the same J and f functions described in the proof of Lemma 2.6. Choose a point β' (whose existence is guaranteed by Lemma 2.6) such that

$$\mu(0) - \mu(\beta') = \lambda > 0$$

say. By the proof of Lemma 2.4, $\mu_N(\beta') \to \mu(\beta')$, so there is N' such that for all $N \geqslant N'$,

$$\mu(0) - \mu_N(\beta') > \lambda/2 \quad .$$

This fact, together with

$$D_N(\beta') = D_N(0) + N(\mu_N(\beta_N') - \mu(0)) + O_p(\sqrt{N}),$$

imply that given $\varepsilon > 0$ there is N'' such that

(2.23) $$P\{ D_N(0) - D_N(\beta') \geq N\lambda/4\} \geq 1 - \varepsilon \qquad \text{for all } N \geq N''.$$

We consider the neighborhood about 0 defined by

$$O_o = \{\beta: \ |\beta| < \frac{\lambda}{8B_J B_c}\} \quad .$$

By Lemma 2.5

(2.24) $$\inf_{\beta \in O_o} D_N(\beta) \geq D_N(0) - N\lambda/8 \qquad \text{w.p.1.}$$

Thus by (2.23)

$$P\{ \inf_{\beta \in O_o} D_N(\beta) \geq D_N(\beta') + N\lambda/8\} \geq 1 - \varepsilon$$

for all $N \geq N''$. Hence $\tilde{\beta}_N \overset{P}{\not\to} 0$. §

2.4 Comments on paper of Stigler

In the next several sections we consider [29]: "Linear functions of order statistics with smooth weight functions," by Stephen Stigler. Our interest in this paper derives from the fact that many of the results in it are used extensively in Sections 2.2 and 2.3 in considering the consistency of Jaeckel's estimator $\tilde{\beta}_N$.

However for our purposes there are several deficiencies in the paper: Theorems 6 and 7 are incorrect (cf. Section 2.5 for a counterexample); in our attempt to patch up Theorem 7, it was also discovered that there is a mistake in Stigler's proof of Theorem 4, on which his later results depend; lastly there are gaps in several of his proofs which possibly deserve some elucidation.

In Sections 2.4-2.7 we state and prove a corrected version of Stigler's Theorem 7, and in the course of the proof we indicate one way of getting around the difficulty involved in his Theorem 4 (at the expense of an extra assumption, however); in these sections we also prove some results which help to fill in the gaps. In the remainder of this section, we outline the problem which Stigler's paper addresses, his notation, and outline his most important results and their interconnections, in order to illuminate our later proofs (which lean heavily on Stigler's proofs).

Let $X_{1n}, X_{2n}, \ldots, X_{nn}$ be independent random variables with (possibly different) cdfs $F_{1n}, F_{2n}, \ldots, F_{nn}$. If we denote the order statistics of this sample as $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, then Stigler is interested in the asymptotic behavior of the statistic

$$(2.25) \qquad S_n = n^{-1} \sum_{i=1}^{n} J(i/(n+1))\, X_{(i)} \quad ,$$

where $J: [0,1] \to R$ is some weight function. By asymptotic behavior we mean the asymptotic normality of (a normalized version of) S_n

and the speed of convergence of $E(S_n)$ to an asymptotic value. The paper is concerned with two set-ups: the X^s are iid, so $F_{kn} = F$ for all k and n, or the more general non-iid case. In both set-ups Stigler's results deal with the interplay between assumptions about the cdfs (the existence or lack thereof of second moments) and assumptions about the weight function J (whether it "trims" or not).

Theorem 2 is the basic asymptotic normality result for S_n assuming the iid case and that F has a finite second moment. Theorem 4 is the basic result containing information on how fast $E(S_n)$ converges to $\mu(J,F) \equiv \int_0^1 J(u)\, F^{-1}(u)\, du$, also assuming the iid case and the existence of the second moment. The proof of normality basically involves an application of Hajek's projection lemma to show S_n is "equivalent" to a sequence of random variables for which the standard central limit theorem applies. On the other hand the proof of Theorem 4 basically involves an application of dominated convergence -- unfortunately the purported dominating function does not dominate. Theorem 5 then combines the conclusions of Theorems 2 and 4, only the assumption of a second moment is replaced by a much weaker tail condition while the extra assumption that J trims (i.e. $J(u) = 0$ for $u \leqslant \alpha$ and $u \geqslant 1-\alpha$) is added.

Stigler indicates in the proof of Theorem 5 how in general this new assumption can be used at the places where dominated convergence was invoked in the proofs of Theorems 2 and 4 to replace the second moment assumption -- it involves using the new assumption to bound

certain binomial tail probabilities. Lastly Theorem 6 is the extension of Theorems 2 and 4 to the non-iid case, and Theorem 7 extends Theorem 5 to the non-iid case. The proofs on the whole are pretty similar to those in the iid case, only now Lindeberg-Feller replaces the regular central limit theorem, and certain random variables, useful in bounding different quantities, which were binomial are now generalized binomial random variables. Theorems 6 and 7 both contain the rate of convergence result for $E(S_n)$ which is false. The proof to Theorem 6 outlines the changes to the proof of Theorem 2 necessary to prove asymptotic normality in the non-iid case, but no real proof is given to the (false) extension of Theorem 4 -- just that it follows "in an equally straightforward manner." No proof to Theorem 7 is given -- just that it follows from Theorem 6 as Theorem 5 followed from Theorems 2 and 4.

Our proof of a corrected version of Theorem 7 will be very similar to Stigler's development. We will begin by proving the corrected version of Theorem 6: this will be obtained by mimicking the proof Stigler gives (pp. 684-686) for Theorem 4 -- only assuming the non-iid case -- and by using the comments in Stigler's proof of Theorem 6, and by circumventing the incorrect dominating argument by assuming that J trims (in addition to the assumption of second moments). Then the step from this corrected Theorem 6 to the corrected Theorem 7 will invoke the method described in Stigler's proof of Theorem 5; in Section 2.7 we show how to obtain certain inequalities for the generalized binomial distribution necessary for this step.

2.5 Counterexamples to Stigler

In his Theorem 6 Stigler asserts that $n^{\frac{1}{2}}(E(S_n) - \mu(J,\bar{F})) \to 0$ as $n\to\infty$, where

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} F_{kn}(x) = \bar{F}(x).$$

The following example, which satisfies all of the regularity conditions of Theorem 6, shows that in general this result is not true.

Counterexample. Let $X_1, X_2, \ldots$ be iid $U(0,1)$ random variables, and define the triangular array

$$Z_{kn} = X_k + n^{-\frac{1}{4}} \qquad k=1,2,\ldots,n. \tag{2.26}$$

Let F denote the cdf of the X^s and F_{kn} that of Z_{kn}. To satisfy the conditions of the theorem we need also a cdf G with finite second moment and some finite constant M such that $F_{kn}(y) \leqslant G(y)$ if $y \leqslant -M$ and $F_{kn}(y) \geqslant G(y)$ if $y \geqslant M$. For the cdf G in this case take the cdf of a $U(-3,3)$ random variable, and take $M=2$. Then $\{F_{kn}\}$ and G satisfy the requirements. Furthermore it is clear that

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} F_{kn}(x) = F(x), \tag{2.27}$$

so the limit on the left hand side exists as is required by the theorem. In the terminology of the theorem then, $\bar{F}(x) = F(x)$. Also it is clear that

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \left[F_{kn}(\min(x,y)) - F_{kn}(x)\, F_{kn}(y)\right]$$

exists, thus satisfying the requirements of the theorem. We define

$$S_n = n^{-1} \sum_{i=1}^{n} J(i/(n+1))\, X_{(i)} \tag{2.28}$$

$$S_n^* = n^{-1} \sum_{i=1}^{n} J(i/(n+1))\, Z_{(i)} \quad ,$$

where $Z_{(i)}$ is the i^{th} ordered value among $\{Z_{kn}\}_{k=1}^{n}$, and $J(u)$ is any nice function, specifically take $J(u) = 1$. Clearly J satisfies all of the regularity conditions. Then

$$S_n^* = S_n + n^{-\frac{1}{4}} \quad . \tag{2.29}$$

According to the last remarks of Theorem 6

$$\text{(i)} \quad n^{\frac{1}{2}} \left[E(S_n) - \mu(F)\right] \to 0 \tag{2.30}$$

$$\text{(ii)} \quad n^{\frac{1}{2}} \left[E(S_n^*) - \mu(F)\right] \to 0$$

as $n \to \infty$, where

$$\mu(F) = \int_0^1 J(u)\, F^{-1}(u)\, du.$$

However using (2.29) in (ii) we have $n^{\frac{1}{2}} \left[E(S_n) + n^{\frac{1}{4}} - \mu(F)\right] \to 0$ as $n \to \infty$, which contradicts (i). §

As we will show in the next section, a simple and natural modification of the offending statement makes the rate result obtain in Theorem 6 (and also in Theorem 7). In contrast, the error associated with Theorem 4 relates to the proof of the result and not its statement. It may very well be that the result stated in Stigler's Theorem 4 is correct. However we were not able to patch up the proof without introducting an additional assumption.

The error in the proof of Theorem 4 is located in the first

sentence of the last paragraph of p. 685 of [29]: namely the statement that the supremum of $H_n^*(u;x)$ occurs at $u=F(x)$ is not in general true. For suppose x is chosen such that $F(x) = \frac{j+\eta}{n+1}$ for some nonnegative integer j and $0<\eta<1$ (note the strict inequalities). Obviously this is the case for a.e. **x** since F has a positive density. We now compute $H_n^*(F(x)^-;x)$ and $H_n^*(F(x);x)$. (To shorten notation we will usually just write $H_n^*(u)$ for $H_n^*(u;x)$.) For $u<F(x)$,

$$(2.31) \qquad H_n^*(u;x) = H_n(u;x) = n^{-\frac{1}{2}} \cdot \sum_{i\leqslant(n+1)u} P(X_{(i)}>x) .$$

For convenience we denote $P(X_{(i)}>x) = p_i$ (suppressing the dependence on n and x). We consider $u\varepsilon[j/(n+1), F(x))$ (in this interval it turns out H_n^* is constant). For such u, $[(n+1)u] = j$, where the brackets denote the greatest integer function. Thus

$$H_n^*(u) = n^{-\frac{1}{2}} \sum_{i=1}^{j} p_i ,$$

and hence

$$(2.32) \qquad H_n^*(F(x)^-) \equiv \lim_{u\uparrow F(x)} H_n^*(u) = n^{-\frac{1}{2}} \sum_{i=1}^{j} p_i .$$

Now we consider $u=F(x)$. Note that

$$(2.33) \qquad H_n^*(F(x)) = H_n(F(x)) - a_n(F(x)) .$$

But $H_n(F(x)) = n^{\frac{1}{2}}(1 - F(x)) - n^{-\frac{1}{2}} \cdot \sum_{i>(n+1)F(x)} p_i$. By calculations

done after equation (12) of p. 684,

(2.34) $$\sum_{i=1}^{n} p_i = n - nF(x) \quad ,$$

implying

$$\sum_{i>(n+1)F(x)} p_i = \sum_{i=j+1}^{n} p_i = n - nF(x) - \sum_{i=1}^{j} p_i,$$

so

(2.35) $$H_n(F(x)) = H_n^*(F(x)^-) \quad .$$

But $a_n(F(x)) = n^{-\frac{1}{2}}\,[(n+1)F(x)] - n^{\frac{1}{2}}F(x) = n^{-\frac{1}{2}}j - n^{\frac{1}{2}}((j+\eta)/(n+1))$,

so

(2.36) $$a_n(F(x)) = \frac{j - n\eta}{\sqrt{n}\,(n+1)} \quad .$$

Combining (2.33), (2.35), and (2.36) we obtain

(2.37) $$H_n^*(F(x)) = H_n^*(F(x)^-) - \left(\frac{j - n\eta}{\sqrt{n}\,(n+1)}\right) \quad .$$

If $\eta<j/n$, then $j-n\eta>0$, implying $H_n^*(F(x)) < H_n^*(F(x)^-)$, in turn implying that $\sup_u H_n^*(u;x)$ does not necessarily occur at $u=F(x)$. Indeed for any x such that

$$n\{(n+1)F(x) - [(n+1)F(x)]\} < [(n+1)F(x)],$$ the supremum will not be at $u=F(x)$. However, from the fact that H_n^* is non-decreasing for $u<F(x)$ and non-increasing for $u\geq F(x)$ and the above

calculations, we can say

$$\sup_u H_n^*(u;x) \leqslant H_n^*(F(x);x) + |a_n(F(x))| \ . \tag{2.38}$$

The main difficulty that the correct statement (2.38) causes in the proof is not in showing that

$$\left|\int_0^1 \{J(u) - J(F(x))\}\, dH_n^*(u;x)\right| \to 0,$$

but instead in bounding this sequence by an integrable function so that dominated convergence can be invoked.

2.6 Proofs of corrected results

We will now prove a modified version of Theorem 6:

Theorem 6* Suppose that for some distribution function $G(y)$ of a random variable Y with $E(Y^2)<\infty$, it is true that whenever $y \leqslant -M$, $F_{kn}(y) \leqslant G(y)$, and whenever $y \geqslant M$, $F_{kn}(y) \geqslant G(y)$, where M is some finite constant. Assumethat for a.e. x,y with respect to Lebesgue measure the following limits exist:

$$\lim_{n\to\infty} \bar{F}_n(x) \equiv \bar{F}(x), \text{ where } \bar{F}_n(x) = n^{-1} \cdot \sum_{k=1}^{n} F_{kn}(x) \quad ;$$

and

$$\lim_{n\to\infty} n^{-1} \cdot \sum_{k=1}^{n} \{F_{kn}(\min(x,y)) - F_{kn}(x)\, F_{kn}(y)\} \equiv K(x,y) \ .$$

Then if $J(u)$ is bounded and continuous a.e. $\bar{F}^{-1}$,

$$n\sigma^2(S_n) \to \sigma^2(J,\bar{F},K) \quad \text{(given below) as } n\to\infty;$$

and if $\sigma^2(J,\bar{F},K)>0$, then

$$\frac{S_n - E(S_n)}{\sigma(S_n)} \xrightarrow{D} N(0,1) \quad \text{as } n\to\infty .$$

Here

$$\sigma^2(J,\bar{F},K) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\bar{F}(x))\, J(\bar{F}(y))\, K(x,y)\, dxdy .$$

If in addition J satisfies a Hölder condition with $\gamma>\frac{1}{2}$ (cf. Section 2.1) (except possibly at a finite set of points $t_1,\ldots,t_p$ with $\bar{F}_n^{-1}$ measure 0 for each n) and $J(u)=0$ for $u\leq\alpha$ and $u\geq 1-\alpha$, and if

$$\int\{G(y)\ (1-G(y))\}^{\frac{1}{2}}\, dy$$

is finite, then

$$n^{\frac{1}{2}}\ (E(S_n) - \mu(J,\bar{F}_n)) \to 0, \qquad \text{where} \tag{2.39}$$

$$\mu(J,\bar{F}_n) = \int_0^1 J(u)\ \bar{F}_n^{-1}(u)\ du. \quad \S$$

<u>Proof</u> The proof of the asymptotic normality is just Stigler's proof on page 689 (cf. Section 2.7 however for elucidation on the use of Chebychev's inequality for the generalized binomial distribution). The task of this proof is to show that (2.39) obtains by appropriately modifying Stigler's proof of Theorem 4. (The following should be read in conjunction with that proof.) Without loss of generality we assume 0 is a median of $\bar{F}$. As in his proof integration by parts yields

$$(2.40)\qquad E(S_n) = \int_0^\infty \{n^{-1}\sum_{i=1}^{n} J(\tfrac{i}{n+1})\, P(X_{(i)} > x)\}\, dx - \int_{-\infty}^{0} \{n^{-1}\sum_{i=1}^{n} J(\tfrac{i}{n+1})\, P(X_{(i)} \leqslant x)\}\, dx \quad .$$

We will handle only the first integral, since the result for the second follows in a similar manner. We define

$$I_n = n^{\frac{1}{2}} \int_0^\infty \{n^{-1}\sum_{i=1}^{n} J(\tfrac{i}{n+1})\, P(X_{(i)} > x) - \int_{\overline{F}_n(x)}^{1} J(u)\, du\}\, dx$$

$$= \int_0^\infty \int_0^1 J(u)\, d\overline{H}_n(u;x)\, dx \quad ,$$

where

$$\overline{H}_n(u;x) = \begin{cases} n^{-\frac{1}{2}}(1-u) - n^{-\frac{1}{2}} \cdot \sum_{i>(n+1)u} P(X_{(i)} > x) & u \geqslant \overline{F}_n(x) \\ n^{-\frac{1}{2}} \cdot \sum_{i \leqslant (n+1)u} P(X_{(i)} > x) & u < \overline{F}_n(x) \quad . \end{cases}$$

Define

$$a_n(u) = n^{-\frac{1}{2}}\left[(n+1)u\right] - n^{\frac{1}{2}}u \qquad (\text{with } a_n(1) \equiv 0),$$

and

$$\overline{H}_n^*(u;x) = \overline{H}_n(u;x) - a_n(u)\, I_{\{u \geqslant \overline{F}_n(x)\}}(u) \quad .$$

Then $$I_n = I_{1n} + I_{2n} \quad , \text{ where}$$

$$I_{1n} = \int_0^\infty \int_0^1 \{J(u) - J(\overline{F}_n(x))\}\, d\overline{H}_n^*(u;x)\, dx \quad ,$$

and

$$I_{2n} = \int_0^\infty \{\int_{\overline{F}_n(x)}^{1} J(u)\, da_n(u) + J(\overline{F}_n(x))\, a_n(\overline{F}_n(x))\}\, dx \quad .$$

We shall show I_{1n} and $I_{2n} \to 0$ as $n \to \infty$ by the dominated convergence theorem.

We note that $\overline{H}_n^*(u)$ is monotone non-decreasing for $u < \overline{F}_n(x)$, and monotone non-increasing for $u \geqslant \overline{F}_n(x)$; by the reasoning leading

to (2.38)

$$\sup_u \overline{H}_n^*(u) \leq \overline{H}_n^*(\overline{F}_n(x)) + |a_n(\overline{F}_n(x))| \quad . \tag{2.41}$$

We also note that now the X_i's are not iid F, but instead $X_{kn} \sim F_{kn}$, $k = 1,2,\ldots,n$; thus $P(X_{(i)} > x)$ is no longer the (lower) tail probability of a binomial random variable, but is the tail probability of the generalized binomial random variable $Y_n = \#\{X^s \leq x\}$ with parameter $\underset{\sim}{p}_n^T = (F_{1n}(x), \ldots, F_{nn}(x))$, and mean $n\overline{F}_n(x)$. Consider a fixed $x \neq \overline{F}^{-1}(t_1), \ldots, \overline{F}^{-1}(t_p)$. Note that for all n sufficiently large, $\min(|\overline{F}_n(x) - t_1|, \ldots, |\overline{F}_n(x) - t_p|) \geq n^{-1/8}$.

Defining $\overline{A}_n = \{u\colon |u - \overline{F}_n(x)| \leq n^{-1/8}$ and $0 \leq u \leq 1\}$, it follows by Chebychev's inequality for fourth powers of a generalized binomial random variable (cf. Section 2.7 also) and the boundedness of J that

$$\left| \int_0^1 \{J(u) - J(\overline{F}_n(x))\} d\overline{H}_n^*(u;x) \right| \tag{2.42}$$

$$= \left| \int_{\overline{A}_n} \{ J(u) - J(\overline{F}_n(x))\} \, d\overline{H}_n^*(u;x) \right| + o(1)$$

$$\leq \sup_{u \in \overline{A}_n} |J(u) - J(\overline{F}_n(x))| \cdot |d\overline{H}_n^*(u;x)| + o(1) \quad ,$$

where $|d\overline{H}_n^*|$ is the total variation of the measure $\overline{H}_n^*$. Since $\overline{H}_n^*$ is monotone increasing and then monotone decreasing,

$$|d\overline{H}_n^*| \leq 2 \cdot \sup_u \overline{H}_n^*(u;x) \quad .$$

Thus the right hand side of (2.42) is less than or equal to

$$2\cdot\sup_{u\,\varepsilon\,\overline{A}_n}|J(u) - J(\overline{F}_n(x))|\cdot \sup_u \overline{H}_n^{\,*}(u;x) + o(1).$$

Since the length of $\overline{A}_n$ goes to 0, $\sup_{u\,\varepsilon\,\overline{A}_n}|J(u) - J(\overline{F}_n(x))| \to 0$ as $n \to \infty$. Thus to show $I_{1n} \to 0$ we need only show that $\sup_u \overline{H}_n^{\,*}(u;x)$ is bounded, and that the sequence of functions

$$\left| \int_0^1 \{J(u) - J(\overline{F}_n(x))\}\, d\overline{H}_n^{\,*}(u;x)\right|$$

on $x \,\varepsilon\, [0,\infty)$ is bounded by an integrable function.

As in Stigler's proof

$$(2.43) \qquad \overline{H}_n^{\,*}(\overline{F}_n(x);x) = n^{-\frac{1}{2}}\, E\,\{\max(n\overline{F}_n(x) - [(n+1)\overline{F}_n(x)],\ n\overline{F}_n(x) - Y_n)\}\ .$$

Now 0 is a median of $\overline{F}$, but for $x \geqslant 0$ we don't know that $\overline{F}_n(x) \geqslant .5$. However, since $|n\overline{F}_n(x) - [(n+1)\overline{F}_n(x)]| \leqslant 2$ certainly,

$$(2.44) \qquad \max\,(n\overline{F}_n(x) - [(n+1)\overline{F}_n(x)],\ n\overline{F}_n(x) - Y_n) \leqslant |n\overline{F}_n(x) - Y_n| + 2.$$

Then by Stigler's reasoning

$$\overline{H}_n^*\,(\overline{F}_n(x)) \leqslant \sqrt{\overline{F}_n(x)\,(1 - \overline{F}_n(x))} + 2/\sqrt{n}\ .$$

Since it is easy to show that $|a_n(F_n(x))| \leqslant 1/\sqrt{n}$, this result combined with (2.41) yields (for all n sufficiently large)

$$\begin{aligned}\sup_u \overline{H}_n^*\,(u;x) &\leqslant \overline{H}_n^*(\overline{F}_n(x)) + |a_n(\overline{F}_n(x))| \\ &\leqslant \sqrt{\overline{F}_n(x)\,(1 - \overline{F}_n(x))} + 1 \\ &\leqslant 2\ ,\end{aligned}$$

implying the boundedness of $\sup_u \overline{H}^*_n(u;x)$.

Denoting $\rho_n(x) = \int_0^1 \{J(u) - J(\overline{F}_n(x))\}\, d\overline{H}^*_n(u;x)$,

we now need to find a function $\rho(x)$, integrable on $x \varepsilon [0,\infty)$, such that $|\rho_n(x)| \leqslant \rho(x)$ for all n sufficiently large and all $x \varepsilon [0,\infty)$. We write $[0,\infty) = L_1 \cup L_2 = [0,\overline{F}^{-1}(1-\frac{\alpha}{4})) \cup [\overline{F}^{-1}(1-\frac{\alpha}{4}),\infty)$. For $x \varepsilon L_1$ it is easy to define $\rho(x)$:

$$
\begin{aligned}
\rho_n(x) &\leqslant \int_0^1 |J(u) - J(\overline{F}_n(x))|\ |d\overline{H}^*_n| \\
&\leqslant \int_0^1 2B_J \{2\sqrt{\overline{F}_n(x)\,(1 - \overline{F}_n(x))} + 2\}\, du \\
&\leqslant \int_0^1 2B_J \{2\cdot\tfrac{1}{2} + 2\}\, du \leqslant 6B_J .
\end{aligned}
$$

Thus we set $\rho(x) = 6B_J$ for $x \varepsilon L_1$.

$x \varepsilon L_2$: Let $x_0 = \overline{F}^{-1}(1 - \frac{\alpha}{4})$; then by assumption $\overline{F}_n(x_0) \to 1 - \frac{\alpha}{4}$,

so there is N_1 such that for $n \geqslant N_1$ $\overline{F}_n(x_0) \geqslant 1 - \frac{\alpha}{2}$. Since $\overline{F}_n$ is monotone this implies $\overline{F}_n(x) \geqslant 1 - \frac{\alpha}{2}$ for all $x \varepsilon L_2$ if $n \geqslant N_1$. But $J(u) = 0$ for $u \geqslant 1-\alpha$, so for $x \varepsilon L_2$ and $n \geqslant N_1$ $J(\overline{F}_n(x)) = 0$, implying

$$\rho_n(x) = \int_0^{1-\alpha} J(u)\, d\overline{H}^*_n(u;x) \quad \text{for } x \varepsilon L_2,\ n \geqslant N_1 .$$

Since $\overline{H}^*_n$ is monotone increasing for $u \leqslant \overline{F}_n(x)$, and since $\overline{F}_n(x) \geqslant 1 - \frac{\alpha}{2}$ we obtain

$$|\rho_n(x)| \leqslant B_J\, \overline{H}{}^*_n\,(1-\alpha;x) \qquad \text{for } x\varepsilon L_2,\ n \geq N_1$$

$$= B_J\, \overline{H}_n(1-\alpha;x) \qquad ,$$

since $u = 1-\alpha < \overline{F}_n(x)$ and $\overline{H}{}^*_n = \overline{H}_n$ for $u<\overline{F}_n(x)$. Thus

$$|\rho_n(x)| \leqslant B_J\, n^{-\frac{1}{2}} \cdot \sum_{i\leqslant(n+1)(1-\alpha)} P(X_{(i)}>x) \quad . \tag{2.45}$$

Any index i in this summation satisfies $i\leqslant(n+1)(1-\alpha)$, so for all $n \geqslant 4(1-\alpha)/\alpha$ say,

$$|\mu_{Y_n} - i| \geqslant \frac{n\alpha}{4} \quad , \quad \text{where } P(X_{(i)}>x) = P(Y_n<i)$$

and $E(Y_n) = \mu_{Y_n} = n\overline{F}_n(x) \geqslant n(1 - \frac{\alpha}{2})$. By Chebychev's inequality for fourth moments

$$P\{|Y_n - \mu_{Y_n}| \geqslant \varepsilon\} \quad \leqslant \quad \frac{\mu_4(Y_n)}{\varepsilon^4} \quad , \text{ where}$$

$\mu_4(Y_n)$ is the 4^{th} central moment of Y_n. Letting $\varepsilon = \frac{n\alpha}{4}$ and using the results of Section 2.7, we have for any i in the index set

$$P(X_{(i)}>x) \quad \leqslant \quad \frac{3n^2\, \overline{F}{}^2_n\,(x)\,\{1-\overline{F}_n(x)\}^2 + n\overline{F}_n(x)\,\{1-\overline{F}_n(x)\}}{\left(\frac{n\alpha}{4}\right)^4}$$

$$\leqslant \frac{K_\alpha\, \overline{F}_n(x)\,\{1-\overline{F}_n(x)\}}{n^2} \quad , \quad \text{where} \quad K_\alpha = \frac{6\cdot 4^4}{\alpha^4} \quad .$$

Thus

$$|\rho_n(x)| \quad \leqslant \quad B_J\, n^{-\frac{1}{2}} \cdot \sum_{i\leqslant(n+1)(1-\alpha)} P(X_{(i)}>x)$$

$$\leqslant \frac{K_\alpha B_J \, \overline{F}_n(x) \, (1-\overline{F}_n(x))}{n^{3/2}}$$

$$\leqslant K_\alpha B_J \, (1-\overline{F}_n(x))$$

for $x\varepsilon L_2$ and all n sufficiently large. Thus for $x\varepsilon L_2$ set

$$\rho(x) = \begin{cases} K_\alpha B_J & x<M \\ K_\alpha B_J \, (1-G(x)) & x\geqslant M \end{cases} .$$

Then for $x\varepsilon[0,\infty)$ $|\rho_n(x)| \leqslant \rho(x)$ and clearly $\int_0^\infty \rho(x)\, dx <\infty$

since $E(Y^2)<\infty$ implies $\int_0^\infty (1-G(x))dx < \infty$ (indeed it implies that

$\int_0^\infty x(1-G(x))\, dx < \infty$ by a simple Fubini argument). Thus we can conclude $I_{1n} \to 0$ as $n\to\infty$. Stigler's proof that $I_{2n} \to 0$ in Theorem 4 is simply adapted to show $I_{2n} \to 0$ in the non-iid case by replacing F(x) by $\overline{F}_n(x)$ and noting that his results still obtain. The proof of Theorem 6* is complete. §

Theorem 7* If the moment conditions on G of Theorem 6* are replaced by the condition that for some $\varepsilon_1>0$, $\lim_{x\to\infty} x^{\varepsilon_1}(1-G(x)-G(-x)) = 0$, and we continue to assume the conditions on J in Theorem 6*, including that J(u) = 0 for $u\leqslant\alpha$ and $u\geqslant 1-\alpha$, then the conclusions of Theorem 6* continue to hold. §

As we have mentioned, the proof of this modification of Theorem 6* utilizes exactly the techniques of Stigler's modification of his proofs of his Theorems 2 and 4 to obtain his Theorem 5: namely one finds bounds

for certain (generalized) binomial tail probabilities which allow the continued use of the dominated convergence theorem under the modified assumptions. We will not prove the result in detail. Section 2.7, however, proves that certain moments of the generalized binomial distribution behave nicely, so that bounds for the needed tail probabilities in the non-iid case are analogous to those Stigler uses in proving Theorem 5 in the iid case.

2.7 Miscellaneous results

Throughout this section let Y be a generalized binomial random variable with parameters n and $\underset{\sim}{p}_n^T = (p_1, p_2, \ldots, p_n)$; i.e. $Y = Z_1 + Z_2 + \ldots + Z_n$, where the Z^s are independent Bernoulli random variables, with $E(Z_i) = p_i$. Let X be a (regular) binomial random variable with parameters n and $\bar{p}$, where $\bar{p} = n^{-1}(p_1 + \ldots + p_n)$.

Lemma 2.7 If we denote the 4^{th} central moment of a random variable W by $\mu_4(W)$ (that is, $\mu_4(W) = E\{(W-\mu_W)^4\}$, where $\mu_W = E(W)$), then

$$\mu_4(Y) \leqslant \mu_4(X). \quad \S$$

(In words: a generalized binomial distribution has smaller fourth central moment than that of the "equivalent" regular binomial distribution.)

Proof The proof will make use of the concept of Schur convexity (cf. [25]). Tedious calculation shows that

$$\mu_4(Y) = \sum_{i=1}^{n} (p_i^4 q_i + q_i^4 p_i) + 3 \sum_{i \neq j} p_i q_i p_j q_j \ ,$$

and $$\mu_4(X) = \sum_{i=1}^{n} \bar{p}^4\, \bar{q} + 3 \sum_{i \neq j} \bar{p}^2 \bar{q}^2 \ , \text{ where}$$

$q_i = 1 - p_i$ $(i=1,\ldots,n)$ and $\bar{q} = 1 - \bar{p}$. Let $S = \{\underset{\sim}{p} \varepsilon R^n : 0 \leqslant p_i \leqslant 1\}$, a convex subset of R^n, and consider $f: S \to R$ defined by

$$f(p_1,\ldots,p_n) = \sum_{i=1}^{n} \left[p_i^4(1-p_i) + (1-p_i)^4 p_i\right]$$

$$+ 3 \sum_{i \neq j} p_i(1-p_i)p_j(1-p_j) \ .$$

Clearly f is symmetric in its arguments. Then by result D.a.7.a of Olkin and Marshall (cf. pp. 8-9 of [25]), to show f is Schur <u>concave</u> it suffices to show

$$(2.46) \qquad (f_{(k)}(\underset{\sim}{p}) - f_{(\ell)}(\underset{\sim}{p}))\,(p_k - p_\ell) \leqslant 0 \quad \text{for all } \underset{\sim}{p} \varepsilon S,$$

where $f_{(k)}(\underset{\sim}{p}) = \dfrac{\partial f(\underset{\sim}{p})}{\partial p_k}$. We now show this. A straightforward calculation shows

$$(2.47) \qquad f_{(k)}(\underset{\sim}{p}) = 1 - 14p_k + 36p_k^2 - 24p_k^3 + T(6-12p_k) \ ,$$

where $T \equiv \sum_{i=1}^{n} p_i(1-p_i)$. This in turn yields

$$(f_{(k)}(\underset{\sim}{p}) - f_{(\ell)}(\underset{\sim}{p}))\,(p_k - p_\ell) = -(p_k - p_\ell)^2 \cdot Q,$$

where $Q = 12T + 14 - 36(p_k + p_\ell) + 24(p_k^2 + p_k p_\ell + p_k^2)$. Thus to show that (2.46) obtains it suffices to show $Q \geq 0$. Clearly

$$T \geq p_k(1-p_k) + p_\ell(1-p_\ell) = p_k + p_\ell - p_k^2 - p_\ell^2 \quad , \text{so}$$

$$Q \geq 12(p_k+p_\ell) - 12p_k^2 - 12p_\ell^2 + 14 - 36(p_k+p_\ell) + 24(p_\ell^2+p_kp_\ell+p_k^2)$$

$$= 14 - 24(p_k+p_\ell) + 12(p_k+p_\ell)^2$$

$$= 12\ (p_k+p_\ell-1)^2 + 2 \ > 0 \quad .$$

Hence f is Schur concave on S, so by Theorem D.a.7 of [25], if $\underset{\sim}{x} \prec \underset{\sim}{y}$ on S, then $f(\underset{\sim}{x}) \geqslant f(\underset{\sim}{y})$ (cf. [25] for notation). But it is well known that (p.23 of [25])

$$(\bar{p},\bar{p},\ldots,\bar{p}) \prec (p_1,p_2,\ldots,p_n) \ , \quad \text{so}$$

$$f(\bar{p},\bar{p},\ldots\bar{p}) \geqslant f(p_1,p_2,\ldots,p_n) \ , \text{ implying } \mu_4(X) \geqslant \mu_4(Y). \ \S$$

<u>Lemma 2.8</u> Denote the k^{th} factorial moment of a random variable W by $\mu'_k\ (W)$ (that is, $\mu'_k\ (W) = E(W^{(k)}) \equiv E\ \{W(W-1)\cdots(W-k+1)\}$). Then

$$\mu'_k\ (Y) \leqslant \mu'_k\ (X) \ . \quad \S$$

<u>Proof</u> If W assumes values 0,1,2,... define $\phi(t) = E(t^W)$. Then

$$\mu'_k\ (W) = E(W^{(k)}) = \phi^{(k)}(1), \quad \text{where } \phi^{(k)} \text{ is the } k^{th}$$

derivative. Simple calculations yield $\phi_X(t) = (\bar{q} + \bar{p}t)^n$ and $\phi_X^{(k)}\ (1) = n^{(k)}\ \bar{p}^k$, where $n^{(k)} \equiv n(n-1)\cdots(n-k+1)$. Also

$$\phi_Y(t) = \prod_{i=1}^{n} (q_i+p_it) \quad .$$

To deal with $\phi_Y^{(k)}\ (1)$ we introduce the elementary symmetric functions.

We again denote the column vector $(p_1,\ldots,p_n)^T \varepsilon R^n$ by $\underset{\sim}{p}_n$; let $E_k^n(\underset{\sim}{p}_n)$ denote the k^{th} elementary symmetric function of n arguments.

<u>Claim</u>: $$E(Y^{(k)}) = \phi_Y^{(k)}(1) = k!\, E_k^n(\underset{\sim}{p}_n) .$$

<u>Proof</u>: We will prove this claim by a double induction on k and n with $k \leq n$ ($k=0,1,2,\ldots$ and $n=1,2,\ldots$). For k=0 and arbitrary n,

$$\phi_Y^{(k)}(1) = \phi_Y(1) = 1, \text{ and } E_k^n(\underset{\sim}{p}_n) = 1,$$

verifying the claim. Now suppose the claim has been verified for all (k',n') pairs with $k' \leq k$ and $n' \leq n$. We verify the claim for $k'=k$ and $n'=n+1$:

Denoting $f(t) = q_{n+1} + p_{n+1}\, t$ and $\underset{\sim}{p}_{n+1} = (p_1,\ldots,p_n,p_{n+1})$, we have $\phi_{n+1}(t) = \phi_n(t)\, f(t)$. Thus

$$\phi_{n+1}^{(k)}(t) = \sum_{\ell=0}^{k} \binom{k}{\ell} \phi_n^{(k-\ell)}(t)\, f^{(\ell)}(t) . \tag{2.48}$$

But $f^{(0)}(t) = f(t)$, $f^{(1)}(t) = p_{n+1}$, and $f^{(\ell)}(t) = 0$ for ℓ greater than one. Thus

$$\phi_{n+1}^{(k)}(1) = \phi_n^{(k)}(1) + kp_{n+1}\, \phi_n^{(k-1)}(1)$$

$$= k!\left[E_k^n(\underset{\sim}{p}_n) + p_{n+1}\, E_{k-1}^n(\underset{\sim}{p}_n)\right]$$

But from a simple picture it is clear that

$$E_k^{n+1}(\underset{\sim}{p}_{n+1}) = E_k^n(\underset{\sim}{p}_n) + p_{n+1}\, E_{k-1}^n(\underset{\sim}{p}_n) ;$$

to complete the induction we need to show that for $k=1,2,\ldots$ that

$\phi_k^{(k)}(1) = k!\ E_k^k\ (\underset{\sim}{p}_k)$. But for $\underset{\sim}{p}_k$ $\phi_k(t) = p_1 \cdots p_k t^k + (c_1 t^{k-1} + \ldots + c_k)$, where here $c_1, \ldots, c_k$ are just some constants. Thus $\phi_k^{(k)}(1) = k! p_1 \cdots p_k$ which equals $k! E_k^k\ (\underset{\sim}{p}_k)$. Thus the claim holds.

Since $E(X^{(k)}) = k!\ E_k^n\ (\bar{p}, \ldots, \bar{p})$, to conclude the proof of the lemma we need to show that

$$E_k^n\ (\bar{p}, .., \bar{p}) \geqslant E_k^n\ (p_1, \ldots, p_n) \; . \tag{2.49}$$

One possible approach to this is to use a theorem due to Marcus and Lopes (cf. p.33 of [4]); but a simpler proof is to again use results from Olkin and Marshall about Schur convex functions. By result D.d.3 on p.31 of [25], $E_k^n\ (\underset{\sim}{x})$ is symmetric and Schur concave on $\{\underset{\sim}{x}: \underset{\sim}{x} = (x_1, \ldots, x_n)\}$. Thus $-E_k^n\ (\underset{\sim}{x})$ is symmetric and Schur convex, so by Theorem D.a.7 of Olkin and Marshall we again have

$$-E_k^n\ (\underset{\sim}{x}) \leqslant -E_k^n\ (\underset{\sim}{y}) \quad \text{if } \underset{\sim}{x} \prec \underset{\sim}{y} \text{ in } R^n.$$

Again noting $(\bar{p}, \ldots, \bar{p}) \prec (p_1, \ldots, p_n)$, we obtain $E_k^n\ (\bar{p}, \ldots, \bar{p}) \geqslant E_k^n\ (p_1, \ldots, p_n)$, thus verifying (2.49) and completing the proof. §

CHAPTER 3. ADAPTING JAECKEL'S ESTIMATOR

3.1 Adaptive estimators and the kink family

The ultimate goal of the work on regression problems which we are considering is to find a method of estimation which is asymptotically efficient (in the sense of minimizing the asymptotic variance) and which gives excellent results in small samples for a wide spectrum of error distribtuions. The experience with the location problem seems to indicate that, although the goal may be technically not obtainable, in spirit one may be able to come close, and the methods and insights derived in its pursuit are very useful.

In either the regression or location problem, if one uses L, M, or R estimators, then once can find an asymptotically optimal estimator (under certain conditions) -- assuming one knows the error cdf F. But as was noted in the introduction, the form of the error distribution is seldom known. A reasonable idea for circumventing this obstacle is to somehow estimate the unknown error distribution and proceed from there. An early implementation of this idea to actually construct usable estimators in the location problem was by Jaeckel in [17]. We briefly consider his approach since ours is very closely related. Jaeckel began with a <u>family</u> of estimators for the location parameter: namely α-trimmed means, with the trimming proportion $\alpha \varepsilon [\alpha_0, \alpha_1]$ $(0<\alpha_0<\alpha_1<\frac{1}{2})$. Two characteristics of this family are of note: it is easily parameterized, and for a wide variety of error distributions

there is a member of the family which does reasonably well. However the trimmed means are usually not asymptotically optimal. Also of great importance is the fact that the asymptotic variance $\sigma^2(\alpha)$ of the α-trimmed mean is a fairly simple function of F and f=F'; this means that one has some hope of estimating $\sigma^2(\alpha)$ on the basis of a moderate-sized sample. This is exactly what Jaeckel does: he constructs an estimator $s^2(\alpha)$ of $\sigma^2(\alpha)$. As his adaptive estimator he then chooses that trimmed mean (with $\alpha\varepsilon[\alpha_0,\alpha_1]$) which has the smallest estimated asymptotic variance. He is then able to show under certain conditions that the asymptotic variance of his adaptive trimmed mean for a given error distribtuion is the same as that of the <u>best</u> trimmed mean for that error distribution (with $\alpha\varepsilon[\alpha_0,\alpha_1]$). (Also see Johns [19] for a more ambitious adapting scheme in the location problem.)

The family of estimators for the vector of slope parameters $\underset{\sim}{\beta}$ we wish to consider are the Jaeckel regression estimators (cf. Section 2.1) with monotone score functions J(u) given by

(3.1) $$J_\xi(u) = \begin{cases} -\xi & u<\frac{1}{2}-\xi \\ u-\frac{1}{2} & \frac{1}{2}-\xi\leq u\leq\frac{1}{2}+\xi \\ \xi & u>\frac{1}{2}+\xi \end{cases}$$

with $\xi\varepsilon[0,\frac{1}{2}]$ (see Figure 3). They resemble the "Wilcoxon" scores $J(u) = u-\frac{1}{2}$, except that at $\frac{1}{2}-\xi$ and $\frac{1}{2}+\xi$ they have "kinks," beyond which they are horizontal.

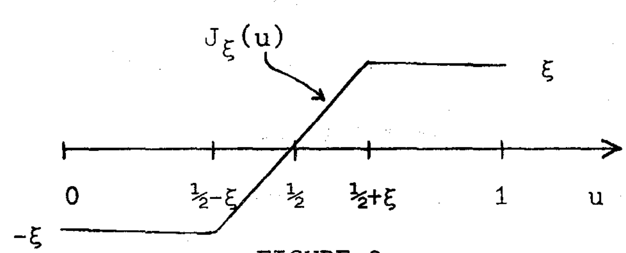

FIGURE 3

We will refer to them simply as kink scores.

A simple calculation using Jaeckel's theorem (see Section 2.1) gives the asymptotic covariance matrix of the estimator with score function $J_\xi(u)$ proportional to (that is, excluding the factor Σ^{-1}) the asymptotic "variance"

$$v(\xi) = \frac{-4\xi^3+3\xi^2}{12\left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt\right]^2} \tag{3.2}$$

We note that the dependence of v on F is confined to the denominator and that it is a relatively nice function of F. This introduces the important question of: why introduce this kink family (or other such simple families) when one might instead think of simply estimating the <u>optimal</u> score function $\phi_f(u) = (-f'/f)(F^{-1}(u))$. In other words, why be concerned with choosing the best estimator from a small family -- which contains optimal scores for only a few distributions -- when one can estimate the optimal score function and base one's estimator on that. The answer lies of course in the virtual impossibility of getting a good pointwise estimator of $-f'/f(F^{-1}(u))$, a quantity involving the second derivative of F. The sample sizes necessary are simply prohibitive. By reducing consideration to the kink family, we have simplified the problem to basically one of estimating the integral of f^2, which we have some hope of doing reasonably well with moderate sample sizes.

To estimate $v(\xi)$ we begin with a preliminary estimate $\tilde{\underset{\sim}{\beta}}_N$ of $\underset{\sim}{\beta}$

satisfying certain conditions (in practice we will just use the Jaeckel estimator with Wilcoxon scores). Then based on the residuals from this preliminary fit we will use standard techniques to construct an estimator $\tilde{f}_N(t)$ of $f(t)$ and an estimator $\tilde{F}_N^{-1}(t)$ of $F^{-1}(t)$. These give us an estimator $\tilde{v}_N(\xi)$ of $v(\xi)$. We then choose as our estimator basically that kink estimator $\underset{\sim}{\beta}_N^*$ which has the smallest estimated asymptotic "variance". More precisely, because of technical reasons, we consider only a <u>finite</u> set of valuse of ξ equally spaced on $[\xi_0,\xi_1]$ $(0<\xi_0<\xi_1<\frac{1}{2})$. As far as results are concerned, this simplification is minor: the estimator obtained in this manner is shown to have asymptotic efficiency (relative to the best kink estimator with $\xi\varepsilon[\xi_0,\frac{1}{2}]$) arbitrarily close to 1. The next section gives the proof of this result.

3.2 Assumptions and Bickel-Rosenblatt result

In this section we outline the assumptions required to establish our results and state the result of Bickel and Rosenblatt [7] which is fundamental to our technique.

The model we are considering is

$$Y_i = \underset{\sim}{c}_i^T \underset{\sim}{\beta} + e_i \qquad i=1,\ldots,N \quad , \tag{3.3}$$

where the e_i are iid random variables with cdf F and density f. We assume we have a preliminary estimate $\tilde{\underset{\sim}{\beta}}_N$ of $\underset{\sim}{\beta}$, satisfying

(i) $\tilde{\beta}_N$ is invariant (cf. p. 1452 of [18])

(ii) $\tilde{\beta}_N = \beta_0 + O_p(N^{-\frac{1}{2}})$, where β_0 is the true value of β.

Because of the invariance we assume without loss of generality that $\beta_0 = 0$, so $\tilde{\beta}_N = O_p(N^{-\frac{1}{2}})$.

As estimates of the unknown errors we use

$$\tilde{e}_j = Y_j - c_j^T \tilde{\beta}_N \qquad j=1,\ldots,N \quad . \tag{3.4}$$

From the assumptions on $\tilde{\beta}_N$, it follows, since we will assume $|c_j| \leq B_c$, that $\tilde{e}_j = e_j + O_p(N^{-\frac{1}{2}})$ uniformly in j. Our estimate of the unknown density f is constructed from these residuals using the density estimator of Bickel and Rosenblatt:

$$\tilde{f}_N(x) = (Nb(N))^{-1} \sum_{j=1}^{N} w((x-\tilde{e}_j)/b(N)) \ , \tag{3.5}$$

where b(N) is a bandwidth going to 0 as $N \to \infty$, and w is a weight function. We also define a density estimator based on the true errors:

$$f_N(x) = (Nb(N))^{-1} \sum_{j=1}^{N} w((x-e_j)/b(N)) \ . \tag{3.6}$$

Our assumptions are:

A1. w is symmetric about 0 with $\int_R w(x)dx = 1$. There is a finite constant A such that w vanishes outside $[-A,A]$; also w is bounded: $|w(x)| \leq B_w$. w has a bounded derivative w' on (-A,A) with $|w'(x)| \leq B_{w'}$. Also, for $(x,x+\delta) \subset (-A,A)$, $w(x+\delta) = w(x) + w'(x)\delta + o(\delta^2)$

uniformly in x.

A2. The density f is continuous, positive, bounded, symmetric, and unimodal. (Without loss of generality we can assume f is symmetric about 0.)

A3. The function $f^{\frac{1}{2}}$ is absolutely continuous and its derivative $\frac{1}{2} f'/f$ is bounded in absolute value. Moreover

$$\int_{[|z|\geq 3]} |z|^{3/2} \left[\log\log|z|\right]^{\frac{1}{2}} \left[|w'(z)| + |w(z)|\right] dz < \infty .$$

A4. The second derivative f" of f exists and is bounded.

A5. $b(N) = o(N^{-2/9})$ and $N^{-\frac{1}{4}} (\log N)^{\frac{1}{2}} (\log\log N)^{\frac{1}{4}} = o(b(N))$ as $N \to \infty$.

A6. $|\underset{\sim}{c}_j| \leqslant B_c$ for j=1,2,...

We note that in our applications we generally will use the "natural" weight function

$$w(t) = \begin{cases} \frac{1}{2} & |t| \leqslant 1 \\ 0 & \text{otherwise} \end{cases} , \tag{3.7}$$

which easily satisfies A1 and the latter part of A3. Assumptions A3, A4, and A5 are of a technical nature which Bickel and Rosenblatt require for their results.

Under Assumptions A1-A6 the following result obtains:

<u>Theorem 3.1</u> (Bickel and Rosenblatt [7])

The quantity

$$b(N)^{-\frac{1}{2}}\left[Nb(N)\int\left[f_N(t)-f(t)\right]^2 a(t)dt - \int f(t)a(t)dt\cdot\int w^2(z)dz\right]$$

is asymptotically distributed $N(0,\ 2\int(\int w(x+y)w(x)dx)^2dy\cdot\int a^2(t)f^2(t)dt\ )$ as $N\to\infty$, where $a(t)$ is a bounded, piecewise smooth integrable function. §

3.3 Preliminary results

In order to estimate the asymptotic variance $v(\xi)$ for different values of ξ, we need to estimate $\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt$. As indicated, to estimate the integrand we use $\tilde{f}_N^2(t)$. To estimate the lower endpoint of integration we use the inverse of the empirical cdf based on the residuals:

$$\tilde{F}_N^{-1}(t) \equiv \inf\{\tilde{e}_{(i)} : i/N\geq t\}.$$

Because of the restriction in Bickel and Rosenblatt that $a(x)$ be integrable, their result cannot be applied without restricting the range of integration $\left[F^{-1}(\frac{1}{2}-\xi),0\right]$. We thus consider the interval of possible ξ values defined as $I = \left[0,\xi_1\right]$, where $\xi_1<\frac{1}{2}$ is arbitrary (in application taken to be close to $\frac{1}{2}$). Then the following result relating

$$\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt$$

and our estimate of this quantity based on the residuals $\{\tilde{e}_j\}$ obtains:

<u>Theorem 3.2</u> Under Assumptions A1-A6

$$\sup_{\xi\in I}\left|\left[\int_{\tilde{F}_N^{-1}(\frac{1}{2}-\xi)}^{0}\tilde{f}_N^2(t)dt\right]^2 - \left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt\right]^2\right| = o_p(N^{-\frac{1}{4}}).\quad §$$

We begin by establishing several lemmas, from which the proof of the theorem will easily follow. For completeness we start with a very useful lemma due to Jaeckel on the behavior of order statistics.

<u>Lemma 3.1</u> (Jaeckel [17]) Let $X_1,\ldots,X_n$ be iid random variables with cdf G. Suppose G is symmetric, has density g, and that there are numbers $\alpha_0>0$, $\varepsilon_0>0$, and $g_0>0$ such that $g(x)\geq g_0$ for all x such that $\alpha_0-\varepsilon_0 \leq G(x) \leq 1-(\alpha_0-\varepsilon_0)$. Then $X_{(i)} - G^{-1}(i/(n+1))$ is $O_p(n^{-\frac{1}{2}})$ uniformly in $i = [\alpha_0 n]+1,\ldots, n-[\alpha_0 n]$. §

We note that if g is unimodal, then Lemma 3.1 is satisfied for any $\alpha_0>0$.

For convenience define $I^* = [F^{-1}(\frac{1}{2}-\xi_1),0]$. Recalling the definition of f_N from (3.6), we have

<u>Lemma 3.2</u> $f_N(x) = O_p(1)$ uniformly for $x\varepsilon I^*$. §

<u>Proof</u> Since w is bounded it suffices to show

$$(3.8)\qquad \sup_{x\varepsilon I^*} \#\{e_j : |x-e_j| \leq Ab(N)\} = O_p(Nb(N)) .$$

Let $x\varepsilon I^*$. Lemma 3.1 implies that given $\varepsilon^*>0$, there exists M such that for all N

$$(3.9)\qquad P\{|e_{(i)}-F^{-1}(i/(N+1))|\leq MN^{-\frac{1}{2}} \text{ for all } i= [\zeta N/4],[\zeta N/4]+1,\ldots, [(1-\tfrac{\zeta}{4})N]\} \geq 1-\varepsilon^* ,$$

where $\zeta\equiv \frac{1}{2}-\xi_1$.

Let Q denote the exceptional set (of probability $<\varepsilon^*$) where these

inequalities may be violated (note that Q does not depend on x). In passing we note that the assumptions about the bandwidth b(N) certainly imply $b(N) = o(1)$ and $N^{-\frac{1}{2}} = o(b(N))$. Now consider $e_{(j)}$ such that $|x-e_{(j)}| \leq Ab(N)$. Then for all N large

$$F^{-1}(\zeta/2) \leq e_{(j)} \leq F^{-1}(1-\frac{\zeta}{2}),$$

since $F^{-1}(\zeta) \leq x \leq 0$ and $Ab(N) \to 0$ as $N \to \infty$. Thus, for $\omega \varepsilon \Omega \sim Q$, we have by (3.9) that for all N sufficiently large

$j \varepsilon \{[\zeta N/4], [\zeta N/4]+1, \ldots, [(1-\frac{\zeta}{4})N]\}$ certainly; so for $\omega \varepsilon \Omega \sim Q$,

$|x-e_{(j)}| \leq Ab(N)$ only if $|x-F^{-1}(j/(N+1))| \leq Ab(N)+MN^{-\frac{1}{2}}$ for N large. But for N large $Ab(N) + MN^{-\frac{1}{2}} \leq 2Ab(N)$; and the number of j^s for which $|x-F^{-1}(j/(N+1))| \leq 2Ab(N)$ is just the number of j^s such that

$$F(x-2Ab(N)) \leq \frac{j}{N+1} \leq F(x+2Ab(N)) \quad ,$$

which is no greater than

$$\begin{aligned} &(N+1)\ (F(x+2Ab(N)) - F(x-2Ab(N))) \\ &\leq 2 + Nf(0) \cdot 4Ab(N) \qquad \text{(unimodality of f)} \\ &\leq O(Nb(N)) \qquad \text{uniformly in x.} \end{aligned}$$

Thus we have shown that

$$\sup_{x \varepsilon I^*} \#\{e_j : |x-e_j| \leq Ab(N)\} = O_p(Nb(N)) \ ,$$

implying the result. §

<u>Lemma 3.3</u> $\sup_{x\varepsilon I^*} |\tilde{f}_N(x) - f_N(x)| = o_p(N^{-\frac{1}{4}})$. §

<u>Proof</u> Let $x\varepsilon I^*$. Then

$$(3.10)\qquad |\tilde{f}_N(x)-f_N(x)| \leqslant (Nb(N))^{-1} \sum_{i=1}^{N} \left| w\left(\frac{x-e_j+(e_j-\tilde{e}_j)}{b(N)}\right) - w\left(\frac{x-e_j}{b(N)}\right)\right| .$$

Let T denote the indices for which the arguments of w on the right hand side (RHS) of (3.10) are both in $[-A,A]$ or both outside of $[-A,A]$; i.e. $T = T_1 \cup T_2$, where

$$T_1 = \{j:\ \frac{x-\tilde{e}_j}{b(N)} \text{ and } \frac{x-e_j}{b(N)}\ \varepsilon[-A,A]\} ,$$

$$T_2 = \{j:\ \frac{x-\tilde{e}_j}{b(N)} \text{ and } \frac{x-e_j}{b(N)}\ \varepsilon[-A,A]^c\} .$$ (Note that

T_1, T_2, and T depend on x and ω.) Since w is zero on $[-A,A]^c$ and $w(y+\delta) - w(y) = \delta w'(y) + o(\delta^2)$ uniformly for y, $y+\delta\varepsilon[-A,A]$, and since $e_j-\tilde{e}_j = O_p(N^{-\frac{1}{2}})$ uniformly in j (independent of x), we obtain that the RHS of (3.10) is less than or equal to

$$(3.11)\qquad (Nb(N))^{-1} \sum_{T_1} \{O_p\left(\frac{N^{-\frac{1}{2}}}{b(N)}\right) \left|w'\left(\frac{x-e_j}{b(N)}\right)\right| + o_p(1/(Nb^2(N)))\} + \frac{2}{Nb(N)} \sum_{T^c} B_w ,$$

where O_p and o_p are uniform for $x\varepsilon I^*$ and in j. Since $|w'|\leqslant B_{w'}=O(1)$,

$$(3.12)\qquad \text{RHS of (3.10)} \leqslant O_p\left(\frac{1}{N^{3/2}b^2(N)}\right) \cdot \#(T_1) + o_p\left(\frac{1}{N^2b^3(N)}\right) \cdot \#(T_1) + O\left(\frac{1}{Nb(N)}\right) \cdot \#(T^c) .$$

Now $\#(T_1)$ is certainly less than or equal to $\#\{e_j : |x-e_j| \leqslant Ab(N)\}$. By (3.8) this latter quantity is $O_p(Nb(N))$ uniformly for $x \varepsilon I^*$. Thus $\sup_{x\varepsilon I^*} \#(T_1) = O_p(Nb(N))$. The assumptions for $b(N)$ imply that for N sufficiently large, $b(N) \geqslant N^{-\frac{1}{4}}(\log N)^{\frac{1}{2}}$, so for all N large the first term on the RHS of (3.12) is $o_p(N^{-\frac{1}{4}})$ uniformly for $x\varepsilon I^*$. Similarly the second term is $o_p(N^{-\frac{1}{2}})$ uniformly for $x\varepsilon I^*$. To evaluate the third term we need to bound $\#(T^c)$ uniformly for $x\varepsilon I^*$.

Let $\varepsilon>0$ be given and consider a fixed N. Then there is M_ε (independent of N) and a subset J_N of Ω, with $P\{J_N\} > 1-\varepsilon$, such that $|\tilde{e}_j - e_j| \leqslant M_\varepsilon N^{-\frac{1}{2}}$ for $\omega\varepsilon J_N$ and $j=1,2,\ldots,N$. Since for N sufficiently large $b(N) \geqslant N^{-\frac{1}{4}}(\log N)^{\frac{1}{2}}$, on J_N we have

$$\left|\frac{\tilde{e}_j - e_j}{b(N)}\right| \leqslant \frac{M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}}$$

for all N large, $j=1,\ldots,N$. Thus for N large and $\omega\varepsilon J_N$

$$\text{(3.13)} \qquad \sup_{x\varepsilon I^*} \#(T^c) \leqslant \sup_{x\varepsilon I^*} \# \left\{ e_j : \frac{x-e_j}{b(N)} \,\varepsilon \left[-A - \frac{M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}}, -A + \frac{M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}} \right] \right.$$
$$\left. \text{or } \frac{x-e_j}{b(N)} \,\varepsilon \left[A - \frac{M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}}, A + \frac{M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}} \right] \right\}.$$

By the same reasoning as in the proof of (3.8), the RHS of (3.13) is

$$O_p\left(\frac{Nb(N)M_\varepsilon}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}}\right).$$ Since $\varepsilon>0$ was arbitrary

we conclude

$$\sup_{x\varepsilon I^*} \#(T^c) = O_p\left(\frac{Nb(N)}{N^{\frac{1}{4}}(\log N)^{\frac{1}{2}}}\right) .$$

Thus the third term on the RHS of (3.12) is $o_p(N^{-\frac{1}{4}})$ uniformly in $x\varepsilon I^*$, and so the RHS of (3.12) is

$$o_p(N^{-\frac{1}{4}}) + o_p(N^{-\frac{1}{2}}) + o_p(N^{-\frac{1}{4}}) = o_p(N^{-\frac{1}{4}}), \text{ uniformly}$$

in $x\varepsilon I^*$. Tracing the inequalities back we have

$$\sup_{x\varepsilon I^*} |\tilde{f}_N(x)-f_N(x)| = o_p(N^{-\frac{1}{4}}) \quad \text{, concluding the proof.§}$$

<u>Corollary 3.1</u> $\tilde{f}_N(x) = O_p(1)$ uniformly for $x\varepsilon I^*$. §

<u>Proof</u> This follows immediately from Lemmas 3.2 and 3.3 . §

<u>Corollary 3.2</u> $\sup_{\xi\varepsilon I} \left|\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} \tilde{f}_N^2(t)dt - \int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f_N^2(t)dt\right| = o_p(N^{-\frac{1}{4}})$.§

<u>Proof</u> Consider a fixed $\xi\varepsilon I$. Note $F^{-1}(\frac{1}{2}-\xi) \geqslant F^{-1}(\frac{1}{2}-\xi_1)$. By Lemma 3.3 $\sup_{t\varepsilon I^*} |\tilde{f}_N(t)-f_N(t)| = o_p(N^{-\frac{1}{4}})$ and by Lemma 3.2 and Corollary 3.1

$$\sup_{t\varepsilon I^*} |\tilde{f}_N(t)+f_N(t)| = O_p(1), \quad \text{implying}$$

$$\sup_{t\varepsilon I^*} |\tilde{f}_N^2(t) - f_N^2(t)| = o_p(N^{-\frac{1}{4}}).$$

From this the result follows. §

<u>Lemma 3.4</u> $\sup_{\xi\varepsilon I} \left|\left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} \tilde{f}_N^2(t)dt\right]^2-\left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f_N^2(t)dt\right]^2\right| = o_p(N^{-\frac{1}{4}})$. §

Proof Let us denote the first integral as $a(\xi)$ and the second as $b(\xi)$, so the quantity of interest is

$$\sup_{\xi\varepsilon I} |a^2(\xi)-b^2(\xi)| = \sup_{\xi\varepsilon I} \{|a(\xi)-b(\xi)|\cdot|a(\xi)+b(\xi)|\} .$$

From Corollary 3.2 we see it suffices to show

$$\sup_{\xi\varepsilon I} |a(\xi)+b(\xi)| = O_p(1) . \tag{3.14}$$

But Lemma 3.2 and Corollary 3.1 imply $f_N^2(t)$ and $\tilde{f}_N^2(t) = O_p(1)$ uniformly for $t\varepsilon I^*$, so $a(\xi)$ and $b(\xi)$ are $O_p(1)$ uniformly for $\xi\varepsilon I$ and so (3.14) follows. §

Lemma 3.5 $\sup_{\xi\varepsilon I} \left|\left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f_N^2(t)dt\right]^2 - \left[\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt\right]^2\right| = o_p(N^{-3/8})$. §

Proof First we show

$$\int_{F^{-1}(\frac{1}{2}-\xi_1)}^{0} [f_N(t)-f(t)]^2 dt = o_p(N^{-3/4}) . \tag{3.15}$$

By the result of Bickel and Rosenblatt (Theorem 3.1),

$$b(N)^{-\frac{1}{2}}\left[Nb(N)\int[f_N(t)-f(t)]^2 a(t)dt - \int f(t)a(t)dt\cdot\int w^2(z)dz\right] = O_p(1). \tag{3.16}$$

Consider $a(t) = I_{[F^{-1}(\frac{1}{2}-\xi_1),0]}(t)$; we note this choice satisfies the assumptions of Bickel and Rosenblatt. Then, since w is bounded and vanishes off $[-A,A]$, we can assert that

$$\int f(t)a(t)dt \cdot \int w^2(z)dz = \text{finite constant.}$$

Also since $b(N) = o(1)$, $b(N)^{\frac{1}{2}} = o(1)$, so that (3.16) implies

$$Nb(N)\int_{F^{-1}(\frac{1}{2}-\xi_1)}^{0} [f_N(t)-f(t)]^2 dt = \text{constant} + o_p(1) = O_p(1),$$

and hence

$$(3.17) \qquad \int_{F^{-1}(\frac{1}{2}-\xi_1)}^{0} [f_N(t)-f(t)]^2 dt = O_p\left(\frac{1}{Nb(N)}\right).$$

Again, for all N large, $b(N) \geq N^{-\frac{1}{4}}(\log N)^{\frac{1}{2}}$, so

$$(Nb(N))^{-1} \leq N^{-3/4}(\log N)^{-\frac{1}{2}} \quad \text{for all N large,}$$

which combined with (3.17) implies (3.15).

We now fix $\xi \varepsilon I$ and let $||g||_\xi \equiv [\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} g^2(t)dt]^{\frac{1}{2}}$, the standard norm on the space of square integrable functions defined on $[F^{-1}(\frac{1}{2}-\xi),0]$. Then the quantity of interest in the lemma (neglecting the sup) is $\left| ||f_N||_\xi^4 - ||f||_\xi^4 \right|$, which is equal to

$$(3.18) \qquad \left| ||f||_\xi - ||f_N||_\xi \right| \cdot \left| ||f||_\xi^3 + ||f||_\xi^2 ||f_N||_\xi + ||f||_\xi ||f_N||_\xi^2 + ||f_N||_\xi^3 \right| .$$

Again using the fact that $f_N(t) = O_p(1)$ uniformly for $t\varepsilon I$, and using $f \leq B_f$, we have $||f_N||_\xi = O_p(1)$ and $||f||_\xi = O(1)$, implying that

$$(3.19) \qquad \left| ||f_N||_\xi^4 - ||f||_\xi^4 \right| = \left| ||f||_\xi - ||f_N||_\xi \right| \cdot O_p(1) .$$

Also, $||a|| = ||(a-b)+b|| \leq ||a-b|| + ||b||$ by Cauchy-Schwarz, implying $||a-b|| \geq ||a|| - ||b||$. Similar reasoning also yields

$$||a-b|| \geq ||b|| - ||a|| ,$$

so $\left| ||a|| - ||b|| \right| \leq ||a-b||$. Thus

(3.20) $$\left| \, ||f||_{\xi} - ||f_N||_{\xi} \right| \leq ||f-f_N||_{\xi} \quad .$$

For any $\xi \varepsilon I$, $||g||_{\xi} \leq ||g||_{\xi_1}$; combining this with (3.20) and (3.19), we obtain

$$\sup_{\xi\varepsilon I} \left| \left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f_N^2(t)\,dt\right]^2 - \left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f^2(t)\,dt\right]^2 \right|$$

$$= ||f-f_N||_{\xi_1} \cdot O_p(1)$$

$$= o_p(N^{-3/8}) \cdot O_p(1) \qquad \text{(by (3.15))}$$

$$= o_p(N^{-3/8}) \ . \quad \S$$

For the random variables $e_1, e_2, \ldots, e_N$, define the inverse empirical cdf F_N^{-1} by

$$F_N^{-1}(t) = \inf\{e_{(i)}: i/N \geq t\}.$$

Lemma 3.6 $$\sup_{\xi\varepsilon I} \left|F^{-1}(\tfrac{1}{2}-\xi) - \tilde{F}_N^{-1}(\tfrac{1}{2}-\xi)\right| = O_p(N^{-\frac{1}{2}}) \ . \ \S$$

Proof It follows easily from Lemma 3.1, since for $\xi\varepsilon I$ $\frac{1}{2}-\xi \geq \zeta$, that

(3.21) $$\sup_{\xi\varepsilon I} \left|F^{-1}(\tfrac{1}{2}-\xi) - F_N^{-1}(\tfrac{1}{2}-\xi)\right| = O_p(N^{-\frac{1}{2}}) \ .$$

Thus we consider the difference between F_N^{-1} and $\tilde{F}_N^{-1}$. Consider a fixed $\xi\varepsilon I$. By our definition of F_N^{-1} and $\tilde{F}_N^{-1}$, we have

$$F_N^{-1}(\tfrac{1}{2}-\xi) = e_{(k)} \text{ or } e_{(k+1)} \ ,$$

depending on whether $(\frac{1}{2}-\xi)N$ is an integer or not, where $k = [(\frac{1}{2}-\xi)N]$ (brackets representing the greatest integer function);

also $\tilde{F}_N^{-1}(\frac{1}{2}-\xi) = \tilde{e}_{(k)}$ or $\tilde{e}_{(k+1)}$. To establish that

$$\sup_{\xi\varepsilon I} |F_N^{-1}(\tfrac{1}{2}-\xi) - \tilde{F}_N^{-1}(\tfrac{1}{2}-\xi)| = 0_p(N^{-\frac{1}{2}}) ,$$

it suffices to show (for example)

$$(3.22) \qquad |e_{(k+1)} - \tilde{e}_{(k)}| = 0_p(N^{-\frac{1}{2}}) \text{ uniformly for } \xi\varepsilon I$$

(the other cases follow similarly). We recall that $\tilde{e}_j = e_j + 0_p(N^{-\frac{1}{2}})$ uniformly in j. This easily implies

$$\tilde{e}_{(j)} = e_{(j)} + 0_p(N^{-\frac{1}{2}}) \text{ uniformly in j.}$$

Hence $e_{(k)} - \tilde{e}_{(k)} = 0_p(N^{-\frac{1}{2}})$ uniformly for $\xi\varepsilon I$. Thus to obtain (3.22) it certainly suffices to show

$$e_{(i+1)} - e_{(i)} = 0_p(N^{-\frac{1}{2}}) \text{ uniformly for}$$
$$i = [\zeta N], [\zeta N]+1, \ldots, [(1-\zeta)N] .$$

But Lemma 3.1 implies

$$\begin{cases} e_{(i)} = F^{-1}(i/(N+1)) + 0_p(N^{-\frac{1}{2}}) \\ e_{(i+1)} = F^{-1}((i+1)/(N+1)) + 0_p(N^{-\frac{1}{2}}) \end{cases} ,$$

uniformly in i in this range. For all N large and all i in this range

$$F^{-1}(\zeta/2) \leqslant F^{-1}(i/(N+1)) \leqslant F^{-1}(1-\zeta/2) \text{ certainly,}$$

so $f(F^{-1}(i/(N+1))) \geqslant f_0 \equiv f(F^{-1}(\zeta/2))$ for all of these i by the unimodality of f. Then by the mean value theorem and unimodality,

$$F^{-1}((i+1)/(N+1)) - F^{-1}(i/(N+1)) \leqslant \frac{1}{(N+1)f_0} = O(N^{-1}).$$

Thus (3.22) obtains, which, combined with (3.21), completes the proof. §

Let us define

$$(3.23)\qquad \begin{cases} a^*(\xi) = \int_{\tilde{F}_N^{-1}(\frac{1}{2}-\xi)}^{0} \tilde{f}_N^2(t)\,dt & \text{and} \\ b^*(\xi) = \int_{F^{-1}(\frac{1}{2}-\xi)}^{0} \tilde{\tilde{f}}_N^2(t)\,dt & . \end{cases}$$

By Lemma 3.6, given $\varepsilon>0$, there is M' such that for all N large, for all $\xi\varepsilon I$, with probability at least $1-\varepsilon/2$ the following inequalities obtain:

$$\begin{cases} \tilde{F}_N^{-1}(\frac{1}{2}-\xi) \geqslant F^{-1}(\frac{1}{2}-\xi_1) & \text{and} \\ |\tilde{F}_N^{-1}(\frac{1}{2}-\xi) - F^{-1}(\frac{1}{2}-\xi)| \leqslant M'N^{-\frac{1}{2}} & . \end{cases}$$

Furthermore, by Corollary 3.1 there exists M'' such that

$$\tilde{f}_N(t) \leqslant M'' \qquad \text{for all } t\varepsilon\left[F^{-1}(\tfrac{1}{2}-\xi_1),0\right]$$

with probability at least $1-\varepsilon/2$. Combining these results we obtain, with probability at least $1-\varepsilon$, for all N large

$$(3.24)\qquad \begin{cases} |a^*(\xi) - b^*(\xi)| \leqslant \frac{M''^2M'}{\sqrt{N}} & \text{for all } \xi\varepsilon I \text{ and} \\ a^*(\xi) \leqslant M''^2M' & \text{for all } \xi\varepsilon I\ . \end{cases}$$

More briefly, $\sup_{\xi\varepsilon I} |a^*(\xi) - b^*(\xi)| = O_p(N^{-\frac{1}{2}})$ and $\sup_{\xi\varepsilon I} a^*(\xi) = O_p(1)$.

We now use these results to prove

<u>Lemma 3.7</u> $$\sup_{\xi\varepsilon I} |a^{*2}(\xi) - b^{*2}(\xi)| = O_p(N^{-\frac{1}{2}}) \text{ . §}$$

<u>Proof</u> From previous work we know $\sup_{\xi\varepsilon I} b^*(\xi) = O_p(1)$. Thus

$$\sup_{\xi\varepsilon I} |a^*(\xi) + b^*(\xi)| = O_p(1).$$

Hence

$$\begin{aligned}\sup_{\xi\varepsilon I} |a^{*2}(\xi) - b^{*2}(\xi)| &= \sup_{\xi\varepsilon I} \left[|a^*(\xi) + b^*(\xi)|\cdot|a^*(\xi) - b^*(\xi)|\right] \\ &= O_p(1) \cdot O_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}}) \text{ . §}\end{aligned}$$

We now return to

<u>Proof of Theorem</u> The quantity of interest is:

$$\begin{aligned}&\sup_{\xi\varepsilon I} \left|\left[\int^0_{\tilde{F}_N^{-1}(\frac{1}{2}-\xi)} \tilde{f}_N^2(t)dt\right]^2 - \left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f^2(t)dt\right]^2\right| \\ &\leq \sup_{\xi\varepsilon I} |a^{*2}(\xi)-b^{*2}(\xi)| \\ &\quad + \sup_{\xi\varepsilon I} \left|\left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} \tilde{f}_N^2(t)dt\right]^2 - \left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f_N^2(t)dt\right]^2\right| \\ &\quad + \sup_{\xi\varepsilon I} \left|\left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f_N^2(t)dt\right]^2 - \left[\int^0_{F^{-1}(\frac{1}{2}-\xi)} f^2(t)dt\right]^2\right| \\ &= O_p(N^{-\frac{1}{2}}) + o_p(N^{-\frac{1}{4}}) + o_p(N^{-3/8}) \\ &= o_p(N^{-\frac{1}{4}}), \quad \text{by Lemmas 3.7,3.4, and 3.5 respectively. §}\end{aligned}$$

3.4 Asymptotics for adaptive estimator

We are now ready to consider the adaptive estimator of $\underset{\sim}{\beta}$ in detail. We define the estimated variance by

$$\tilde{v}_N(\xi) = \frac{-4\xi^3 + 3\xi^2}{12\left[\int_{\tilde{F}_N^{-1}(\frac{1}{2}-\xi)}^{0} \tilde{f}_N^2(t)dt\right]^2} . \tag{3.25}$$

Let an interval $[\xi_0, \frac{1}{2}]$ be given, with $\xi_0 > 0$. Then the following asymptotic result obtains for the adaptive kink estimator:

Theorem 3.3 Let $\varepsilon > 0$ be given. Under Assumptions A1-A6 and the assumptions in the statement of Jaeckel's theorem (see p. 19), there is a $\delta > 0$ such that the adaptive kink estimator $\underset{\sim}{\beta}_N^*$ -- defined as any value of $\underset{\sim}{\beta}$ which minimizes the dispersion $\tilde{D}_N(\underset{\sim}{\beta})$ constructed with score function $J_{\tilde{\xi}}$ (J_ξ defined in equation (3.1)), where $\tilde{\xi}$ is any value of $\xi \varepsilon \Xi \equiv \{\xi_0, \xi_0+\delta, \xi_0+2\delta, \ldots, \xi_1 = \frac{1}{2}-\delta\}$ minimizing $\tilde{v}_N(\xi)$ -- has asymptotic efficiency greater than $1-\varepsilon$, where the asymptotic efficiency is computed with respect to the best kink estimator with $\xi \varepsilon [\xi_0, \frac{1}{2}]$. The grid set Ξ does not depend on the unknown error density function f. §

Proof Let f be any density satisfying the assumptions. We will assume the minimum of $\{v(\xi) : \xi \varepsilon \Xi\}$ is unique, and we will denote the minimizing value $\overline{\xi}$. This is to simplify notation; it is otherwise irrelevant to the proof. We break the proof up into several parts.

(i) For the given $\varepsilon > 0$, there is $\delta > 0$ such that if f satisfies the

conditions of the theorem

$$(3.26)\qquad \frac{\inf\limits_{\xi\varepsilon[\xi_0,\frac{1}{2}]} v(\xi)}{v(\bar{\xi})} > 1-\varepsilon \ .$$

Proof of (i): Denote $\int_{F^{-1}(\frac{1}{2}-\xi)}^{0} f^2(t)dt$ by $I(\xi)$. We first show there is a constant K such that

$$(3.27)\qquad \left|\frac{v'(\xi)}{v(\xi)}\right| \leqslant K \quad \text{for all } \xi\varepsilon[\xi_0,\tfrac{1}{2}], \text{ uniformly in f.}$$

An easy calculation yields

$$(3.28)\qquad \frac{v'}{v}(\xi) = \frac{6-12\xi}{3\xi-4\xi^2} - \frac{2f(F^{-1}(\frac{1}{2}-\xi))}{I(\xi)}$$

The first term is bounded for $\xi\varepsilon[\xi_0,\frac{1}{2}]$ and does not depend on f. The second term (neglecting the constant) is

$$(3.29)\qquad \frac{f(F^{-1}(\frac{1}{2}-\xi))}{\int_{\frac{1}{2}-\xi}^{\frac{1}{2}} f(F^{-1}(u))du} \qquad \text{by change of variables;}$$

for $\frac{1}{2}-\xi\leqslant u\leqslant\frac{1}{2}$, $f(F^{-1}(u)) \geqslant f(F^{-1}(\frac{1}{2}-\xi))$ since F^{-1} is increasing and f is unimodal and symmetric about 0. Thus the integral in the denominator of (3.29) is at least $\xi f(F^{-1}(\frac{1}{2}-\xi))$, implying the quantity of (3.29) is at most ξ^{-1}, which is no larger than ξ_0^{-1}, Thus (3.27) obtains. We return to (3.26). Since v is continuous the infimum of $v(\xi)$ on $[\xi_0,\frac{1}{2}]$ is achieved; label a value at which it is achieved ξ', and suppose $\xi'\varepsilon[\xi_0+r\delta,\xi_0+r\delta+\delta)$ for some r. By the mean value theorem

there exists $\xi^* \varepsilon [\xi_0 + r\delta, \xi']$ such that

$$v(\xi_0+r\delta) - v(\xi') = (\xi' - \xi_0 - r\delta)\cdot v'(\xi^*) \quad ,$$

with $v'(\xi^*) > 0$ (since we can assume $v(\xi_0+r\delta) > v(\xi')$). Since $0 < \xi'-\xi_0-r\delta < \delta$, this implies

$$(3.30) \qquad 1 - \frac{v(\xi')}{v(\xi_0+r\delta)} \leqslant \frac{\delta v'(\xi^*)}{v(\xi_0+r\delta)}$$

$$\leqslant \frac{\delta K\, v(\xi^*)}{v(\xi_0+r\delta)} \quad \text{, by (3.27) .}$$

To conclude our proof of (3.26), since $v(\xi_0+r\delta) \geqslant v(\bar{\xi})$ certainly, it will suffice to determine δ so that $\dfrac{\delta K v(\xi^*)}{v(\xi_0+r\delta)} < \varepsilon$, independent of ξ^* and f. This entails bounding $v(\xi^*)/v(\xi_0+r\delta)$. We consider the function

$$u(x) = v(\xi_0+r\delta+x)/\ v(\xi_0+r\delta) \quad \text{for} \quad x \varepsilon [0,\delta] \quad \text{say.}$$

Then

$$u'(x) = \frac{v'(\xi_0+r\delta+x)}{v(\xi_0+r\delta)} \leqslant \frac{Kv(\xi_0+r\delta+x)}{v(\xi_0+r\delta)} = Ku(x) \quad \text{by}$$

(3.27). Thus u satisfies: $u(0) = 1$ and $u'/u\ (x) \leqslant K$, so $u(x) \leqslant \exp(Kx)$ certainly. Hence $\max\limits_{x\varepsilon[0,\delta]} u(x) \leqslant \exp(K\delta)$, which implies

$$\sup_{x\varepsilon[0,\delta)} \frac{v(\xi_0+r\delta+x)}{v(\xi_0+r\delta)} \leqslant \exp(K\delta) \quad \text{, implying}$$

$v(\xi^*)/\ v(\xi_0+r\delta) \leq \exp(K\delta)$ for any $\xi^* \varepsilon [\xi_0+r\delta, \xi']$ and uniformly in f. In other words

$$\frac{\delta K v(\xi^*)}{v(\xi_0+r\delta)} \leq \delta K \exp(K\delta) \quad , \quad \text{and choosing}$$

$\delta = \varepsilon/2K$ for example makes $\delta K \exp(K\delta) < \varepsilon$ (if ε is small) . This concludes the proof of (3.26).

(ii) For any $\eta>0$ there are N_0 and sets A_N, with $P(A_N)>1-\eta$, such that $N \geq N_0$ implies $v(\bar{\xi})/\ v(\tilde{\xi}_N) = 1$ on A_N.

Proof of (ii): Write $v(\xi)$ in (3.2) as $U(\xi)/V(\xi)$ and $\tilde{v}_N(\xi)$ in (3.25) as $U(\xi)/\tilde{V}_N(\xi)$. Then Theorem 3.2 implies

$$\tilde{V}_N(\xi) = V(\xi) + o_p(N^{-1/4}) \quad \text{uniformly in } \xi \varepsilon [\xi_0, \xi_1]. \tag{3.31}$$

Also there exist constants k_1, k_2, k_3, k_4 such that

$$\left.\begin{array}{r} 0 < k_1 < U(\xi) < k_2 \\ \text{and} \quad 0 < k_3 < V(\xi) < k_4 \end{array}\right. \quad \text{for all } \xi \varepsilon [\xi_0, \xi_1] \ ,$$

implying that $\tilde{v}_N(\xi) = v(\xi) + o_p(N^{-1/4})$ uniformly in $\xi \varepsilon [\xi_0, \xi_1]$.

Thus given $\eta>0$ there are sets A_N, with $P(A_N) > 1-\eta$ such that

$$\sup_{\xi \varepsilon [\xi_0, \xi_1]} |\tilde{v}_N(\xi) - v(\xi)| \leq N^{-1/4} \text{ on } A_N \text{ for all N large.}$$

But for all N sufficiently large,

$$2N^{-1/4} < \min_{\xi \varepsilon \Xi \sim \{\bar{\xi}\}} (v(\xi) - v(\bar{\xi})) \quad , \text{ which}$$

implies that on A_N (N sufficiently large)

$$\tilde{v}_N(\bar{\xi}) \leq \min_{\xi \in \Xi \sim \{\xi\}} \tilde{v}_N(\xi) \ ;$$

this inequality implies $\tilde{\xi}_N = \bar{\xi}$ and thus the result.

Thus, if we define $\bar{\beta}_N$ to be the kink estimator derived from the score function $J_{\bar{\xi}}$, we have $\beta^*_N = \bar{\beta}_N$ on the sets A_N for all N sufficiently large. Since $\eta > 0$ was arbitrary this implies

$$\beta^*_N = \bar{\beta}_N + o_p(N^{-\frac{1}{2}}) \quad \text{and so by Mann-Wald}$$

theory the asymptotic distributions of β^*_N and $\bar{\beta}_N$ are the same. Specifically this implies the asymptotic "variance" of β^*_N is $v(\bar{\xi})$, which combined with (3.26) yields the theorem. §

CHAPTER 4. FUTURE WORK

4.1 Possible extensions

There are a number of topics in the area of robust regression considered in this paper which need further study. Undoubtedly the most important theoretical problem still unanswered is the asymptotic normality of Jaeckel-type estimators based on a non-monotone score function for both the simple linear regression and the general linear regression models. There is every reason to believe that, subject to certain restrictions (as indicated by the counterexample of Section 2.3), a result like Jaeckel's theorem (see Section 2.1) obtains for such estimators. The extension of the consistency result of Section 2.2 to normality does not appear simple however, a situation in contrast say with the case of maximum likelihood estimates. Some work the author has done seems to indicate a plausible approach to the extension, although the technique is very complicated.

On the other hand, the extension of the consistency result to the case of a vector parameter $\underset{\sim}{\beta}$ is very straightforward. Indeed the basic proof of Section 2.2 continues to hold with appropriate modifications (for example, the condition $|c_j| \leqslant B_c$ for the regression constants becomes $|\underset{\sim}{c}_j| \leqslant B_c$ where $|\cdot|$ is now the Euclidean norm; similarly the compactness condition for the set of possible parameter values B^o is now that B^o should be compact in R^q, and so forth).

A further extension of the consistency result would result from weakening the boundedness conditions on the $\{c_i\}$ so as to allow $|c_i| \to \infty$ at a sufficiently slow rate (for example one might use Noether's condition: see [20]). Obviously one should also be able to weaken some of the technical conditions on f (Assumption F2) and the score function J (Assumption J2). These extensions would be, very possibly, at the expense of a much more involved set of proofs. Nothing appears to inherently demand the restrictions we invoked however. The unimodality restriction of Assumption F1 is a different matter, as the counter-example indicates. A most intriguing question is: exactly what sort of conditions does one need to impose on the error distribution to insure correct asymptotic behavior for estimators using non-monotone score functions?

With regard to the results on adapting, there are a number of important extensions that would be very desirable. First, there are obviously a number of alternatives to the kink family which could be used for adapting and whose behavior might lead to better estimators, especially for small samples, than the one we proposed. Second, the implications of extending asymptotic normality to estimators based on non-monotone scores would be very important in adapting, since one could then utilize more flexible families containing non-monotone members. In such a case one could realistically hope to construct an adaptive estimator whose asymptotic efficiency, relative to the Cramer-Rao bound, would be very high across a very large nonparametric

class of error distributions, subject only to regularity conditions. When such a result is achieved, we will have a reasonable understanding of the problem of robust regression.

BIBLIOGRAPHY

[1] Adichie, J.N. (1967). Estimates of regression parameters based on rank tests. Annals of Mathematical Statistics 38, 894-904.

[2] Andrews, D.F. (1974). A robust method for multiple linear regression. Technometrics 16, 523-531.

[3] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). Robust estimates of location. Princeton University Press, Princeton.

[4] Beckenbach, E. and Bellman, R. (1971). Inequalities. Springer-Verlag, New York.

[5] Bickel, P.J. (1973). On some analogues to linear combinations of order statistics in the linear model. Annals of Statistics 1, 597-616.

[6] Bickel, P.J. (1975). One-step Huber estimates in the linear model. Journal of the American Statistical Association 70, 428-434.

[7] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. Annals of Statistics 1, 1071-1095.

[8] Chung, K.L. (1968). A course in probability theory. Harcourt, Brace, & World, New York.

[9] Forsythe, Alan (1972). Robust estimation of straight line regression coefficients by minimizing p^{th} power deviations. Technometrics 14, 159-166.

[10] Hájek, J. and Šidak, Z. (1967). Theory of rank tests. Academic Press, New York.

[11] Hodges, J.L., Jr. and Lehmann, E. (1963). Estimates of location based on rank tests. Annals of Mathematical Statistics 34, 598-611.

[12] Hogg, R. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. Journal of the American Statistical Association 69, 909-923.

[13] Huber, Peter (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics 35, 73-101.

[14] Huber, Peter (1972). Robust statistics: a review. Annals of Mathematical Statistics 43, 1041-1067.

[15] Huber, Peter (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Annals of Statistics 1, 799-821.

[16] Jaeckel, L. (1971). Robust estimation of location: symmetry and asymmetric contamination. Annals of Mathematical Statistics 42, 1020-1034.

[17] Jaeckel, L. (1971). Some flexible estimates of location. Annals of Mathematical Statistics 42, 1540-1552.

[18] Jaeckel, L. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. Annals of Mathematical Statistics 43, 1449-1458.

[19] Johns, M.V., Jr. (1974). Nonparametric estimation of location. Journal of the American Statistical Association 69, 453-460.

[20] Jurečková, J. (1969). Asymptotic linearity of rank test statistic in regression parameter. Annals of Mathematical Statistics 40, 1889-1900.

[21] Jurečková, J. (1971). Nonparametric estimate of regression coefficients. Annals of Mathematical Statistics 42, 1328-1338.

[22] Koul, H.L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. Annals of Mathematical Statistics 40, 1950-1979.

[23] Kraft, C. and van Eeden, C. (1972). Linearized rank estimates and signed-rank estimates of the general linear hypothesis. Annals of Mathematical Statistics 43, 42-57.

[24] Mann, H.B. and Wald, A. (1943). On stochastic limit and order relationships. Annals of Mathematical Statistics 14, 217-226.

[25] Marshall, A. and Olkin, I. (1973). Schur-convex functions. Technical Report 74, Department of Statistics, Stanford University.

[26] Relles, D. Robust regression by modified least squares. Thesis, Yale University.

[27] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association 63, 1379-1389.

[28] Stigler, Stephen (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. Journal of the American Statistical Association 68, 872-879.

[29] Stigler, Stephen (1974). Linear functions of order statistics with smooth weight functions. Annals of Statistics 2, 676-693.

[30] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II, and III. Nederl. Akad. Wetensch. Proc. 53, 386-392, 521-525, 1397-1412.

[31] Yohai, V.J. (1974). Robust estimation in the linear model. Annals of Statistics 2, 562-567.